Methods Research Report

Reliability Testing of the AHRQ EPC Approach to Grading the Strength of Evidence in Comparative Effectiveness Reviews

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-2007-10056-I

Prepared by:
RTI International–University of North Carolina Evidence-based Practice Center
Research Triangle Park, NC

Investigators:
Nancy D. Berkman, Ph.D.
Kathleen N. Lohr, Ph.D.
Laura C. Morgan, M.A.
Emily Richmond, M.P.H
Tzy-Mey Kuo, Ph.D, M.P.H.
Sally Morton, Ph.D.
Meera Viswanathan, Ph.D.
Douglas Kamerow, M.D.
Sue West, Ph.D.
Elizabeth Tant, B.A.

AHRQ Publication No. 12-EHC067-EF
May 2012

This report is based on research conducted by the RTI International–University of North Carolina Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10056-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.
None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Berkman ND, Lohr KN, Morgan LC, Richmond E, Kuo TM, Morton S, Viswanathan M, Kamerow D, West S, Tant E. Reliability Testing of the AHRQ EPC Approach to Grading the Strength of Evidence in Comparative Effectiveness Reviews. Methods Research Report. (Prepared by RTI International–University of North Carolina Evidence-based Practice Center under Contract No. 290-2007-10056-I.) AHRQ Publication No. 12-EHC067-EF. Rockville, MD: Agency for Healthcare Research and Quality. May 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A, M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Task Order Officer
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The authors would like to acknowledge the following individuals who provided assistance during this project.

We thank two RTI International colleagues: Carla Bann, Ph.D., for statistical and psychometric consultation, and Loraine Monroe for technical word processing.

We are very indebted to the assistance from our inter-rater reliability reviewers:

Agency for Healthcare Research and Quality
Suchitra Iyer, Ph.D.
Elisabeth Kato, M.D., M.R.P.

University of Alberta EPC
Donna Dryden, Ph.D.
Lisa Hartling, M.Sc., Ph.D.

ECRI Institute EPC
Meredith Noble, M.S.
Jonathan Treadwell, Ph.D.

University of Minnesota EPC
Mary Butler, Ph.D.
Tatyana Shamliyan, M.D.

Oregon Health and Science University EPC
Susan Norris, M.D., M.Sc., M.P.H.
Kim Peterson, M.S.
Shelley Selph, M.D., M.P.H.
Marian McDonagh, Pharm.D.

University of Ottawa EPC
Mohammed Ansari, M.B.B.S., M.Med.Sc., M.Phil.
Alex Tsertsvadze, M.D., M.Sc.
RTI-University of North Carolina EPC
Stacey Lloyd, M.P.H.
Mike Trisolini, Ph.D.

Tufts University-New England Medical Center EPC
Denish Moorthy, M.B.B.S., M.S.
Winifred Yu, M.S., R.D.

Vanderbilt University EPC
Jeff Andrews, M.D.
Melissa McPheeters, Ph.D., M.P.H.

Southern California-Rand EPC
Alicia Ruelaz, M.D.
Zhen Wang, M.S.

Reliability Testing of the AHRQ EPC Approach to Grading the Strength of Evidence in Comparative Effectiveness Reviews

Structured Abstract

Objectives. This project focused on Agency for Healthcare Research and Quality (AHRQ) methods guidance to its Evidence-based Practice Center (EPC) program on grading the strength of evidence (SOE) related to therapeutic interventions. Our project focused on inter-rater reliability testing of the two main components of the AHRQ approach to grading SOE for specific outcomes: (1) scoring evidence on the four required domains (risk of bias, consistency, directness, and precision), separately for randomized controlled trials (RCTs) and observational studies, and (2) developing an overall SOE grade, given the scores for the individual domains.

Data Sources and Methods. We conducted inter-rater reliability testing using data obtained from two published CERs. We designed 10 exercises (5 positive outcomes [benefits] and 5 harms [adverse effects]); all 10 included RCTs, and 6 of the 10 included 1 or more observational studies.

Eleven pairs of reviewers (22 participants) participated in the exercises. Each reviewer independently completed each of the exercises; subsequently, each pair of reviewers reconciled their independent responses.

We calculated summary statistics to describe agreement among reviewers and their difficulty in making each rating assessment. We used logistic regression analysis to describe the relationship between domain scores and the final SOE grade, both in relation to the specific grade selected and level of agreement among reviewers. We examined the change in independent reviewer ratings following reconciliation among reviewer pairs.

Results. The level of independent reviewer inter-rater agreement for domain scores varied considerably from substantial for RCT risk of bias and directness to slight for observational study risk of bias. Agreement on all other domains was either moderate or fair. Agreement was generally better for RCTs than observational studies and agreement among reconciled reviewer pairs was as good as or better than it was for individual independent reviewers.

Agreement on independent reviewer SOE grades was generally poorer than for domain scores. Overall agreement was slight and it was not appreciably better when limited to the exercises that included only RCTs. Neither agreement on domain scores nor agreement about the level of difficulty in evaluating particular domains predicted the overall SOE grades.

When evidence was limited to RCT studies, better SOE grades of moderate or high were related to RCT domain scores' being considered consistent and precise. The inclusion of observational studies, in addition to RCTs, in an exercise was a strong predictor of a poorer SOE grade — namely, either insufficient or low.

Conclusions. Our findings demonstrate that the conclusions reached by experienced reviewers based on the same evidence can differ greatly, particularly when they are faced with bodies of evidence that do not lend themselves to meta-analysis and they need to rely more heavily on their own judgment. Of particular concern is how to deal with (a) outcomes that are evaluated through

a combination of RCTs and observational studies, (b) outcomes that are evaluated through more than one measure and (c) grading evidence that appears to show no difference.

We conclude that additional methodological guidance is needed, including more details and examples, supported by more training, particularly on how best to evaluate the “thornier” bodies of evidence as discussed above. However, some potential will always exist for disagreement even among experienced reviewers. EPC reviewer teams need to be transparent in how they have conducted this task. This will help to ensure that stakeholders can be confident of their interpretation of the evidence.

Our study provided only a first approximation of reviewers’ rationales for differences in SOE decisions. Additional research is needed to understand gaps in guidance that should be filled, areas of insufficient understanding of the guidance itself and how best to overcome that deficit, and complex decisions that may still need to be left to the review team’s substantive expertise.

Contents

Executive Summary ES-1

Introduction 1
Background 1
AHRQ EPC Approach to Evaluating Strength of Evidence 1
Background: Quality of Evidence Reliability Project Conducted by the GRADE Working Group 3
Study Objectives 3
Methods 7
Information From Recent Comparative Effectiveness Reviews 7
Information Provided to Reviewer Participants 9
Participants 9
Data Collection 10
Data Synthesis and Analysis 10
Agreement on Rating Across Reviewers 10
Agreement on Difficulty Across Reviewers 11
Prediction of Strength of Evidence Grade 11
Subgroup Analyses of Independent Reviewer Assessments 12
Reconciliation Across Reviewer Pairs 13
Qualitative Analyses 14
Results 15
Risk of Bias 16
Randomized Controlled Trials 17
Observational Studies 18
Consistency 18
RCTs 18
Observational Studies 19
Directness 21
RCTs 21
Observational Studies 27
Precision 23
RCTs 23
Observational Studies 24
Strength of Evidence 25
Predictions of Strength of Evidence Grades From Domain Scores 28
Predictions of Level of Agreement in Strength of Evidence Grades From Level of Agreement in Domain Scores 29
Subgroup Analyses of Independent Reviewer Assessments 30
Reconciliation across Reviewer Pairs 31
Discussion 34
Overview of the Project 34
Key Question 1. Principal Findings: Domain Scores 34
Key Question 2. Principal Findings: Strength of Evidence Grades 35

Conclusions 37
Methodological Approach 37
Needed Guidance Enhancements and Future Research 39
References 42

Appendixes

Appendix A. Strength of Evidence Summary Tables Used To Complete the Exercises
Appendix B. Instructions Provided to Reviewers To Complete the Exercises

Tables

Table A. Inter-Rater Reliability Results by Domain and Overall Grade ES-3
Table 1. Required Domains in the AHRQ EPC Approach to Grading Strength of Evidence 2
Table 2. Strength of Evidence Grades and Definitions 3
Table 3. Grading Exercises: Description and Main Characteristics of Evidence 7
Table 4. Illustration of Calculation of Change Scores 14
Table 5. Inter-Rater Reliability Results by Domain and Overall Grade 15
Table 6. Summary of RCT Risk of Bias Domain Scores, by Condition and Outcome 16
Table 7. Summary of Observational Study Risk of Bias Domain Scores, by Condition and Outcome 17
Table 8. Summary of RCT Consistency Domain Scores, by Condition and Outcome 19
Table 9. Summary of Observational Study Consistency Domain Scores, by Condition and Outcome 21
Table 10. Summary of RCT Directness Domain Scores, by Condition and Outcome 22
Table 11. Summary of Observational Study Directness Domain Scores, by Condition and Outcome 23
Table 12. Summary of RCT Precision Domain Scores, by Condition and Outcome 24
Table 13. Summary of Observational Study Precision Domain Scores, by Condition and Outcome 25
Table 14. Strength of Evidence Grades, by Condition and Outcome 26
Table 15. Detail of Reconciled Pair Assessments for Two Exercises With Poor Agreement Across Pairs: Antidepressant Harms: Nausea and Rheumatoid Arthritis Benefit: ACR70 27
Table 16. Prediction of Strength of Evidence Grade by Domain Scores: Odds Ratio From Multinomial Logistic Regression Results 29
Table 17. Prediction of Agreement on Strength of Evidence Grade by Agreement on Domain Score: Odds Ratios From Logistic Regression Results 30
Table 18. Percentage Agreement Among Independent Reviewers, by Condition and Outcome . 30
Table 19. Percentage Agreement Among Independent Reviewers, by Reviewer Academic Training, Experience in Conducting Systematic or Comparative Effectiveness Reviews, and Experience Evaluating Strength of Evidence 31
Table 20. Mean Percentage Change in Domain Scores and the Strength of Evidence Grade Following Reconciliation 32

Figures

Figure 1. Analytic framework for inter-rater reliability in EPC approach to grading strength of evidence 6

Executive Summary

Introduction

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Center (EPC) program, sponsors Comparative Effectiveness Reviews (CERs) and other systematic reviews (SRs). To advance this effort, AHRQ developed methodological guidance encompassing many of the steps of such reviews. We focus here on the guidance on grading the strength of evidence related to therapeutic interventions, which was published in 2010 as definitive guidance for EPCs.

The strength of evidence grade is a key indicator of a review team's level of confidence that the studies included in the review collectively reflect the true effect of an intervention on a health outcome. The AHRQ Methods Guide chapter for grading the strength of evidence instructs reviewers to score the body of evidence in relation to each major outcome and each major comparison in relation to each key question of the review.

Our project focused on inter-rater reliability testing of the two main components of the AHRQ approach to the task of grading strength of evidence for specific outcomes in relation to key questions: (1) scoring evidence on the four required domains (risk of bias, consistency, directness, and precision) and (2) developing an overall strength of evidence grade, given the scores for the individual domains.

Thus, our research focused on key questions concerning the performance of these two main tasks among independent reviews and reconciled pairs of reviewers. Our specific key questions are presented with their corresponding results.

Methods

We conducted the inter-rater reliability testing using data obtained from two published CERs focusing on two distinct medical indications and drug treatments: second-generation antidepressants for the treatment of major depressive disorder (MDD) and disease-modifying antirheumatic drugs (DMARDs) for the treatment of rheumatoid arthritis (RA). From the data in these reviews, the study team designed 10 exercises; all 10 included RCTs, and 6 of the 10 included one or more observational studies. Using the same terminology as is used in the AHRQ Methods Guide chapter, observational studies include nonrandomized trials, cohort, cross-sectional and case-control studies. Also, we provided the risk of bias of individual studies as a quality assessment (good, fair, or poor) because the original reviewers used this metric. For outcomes, we specified five positive outcomes (benefits) and five harms (adverse effects) of the therapies under examination.

Eleven pairs of reviewers (22 participants) participated in the exercises: two from AHRQ and 20 from nine EPCs. To replicate "real world" or "real EPC" practices, we aimed for a dual assessment in which each independent reviewer (at a given organization) was also paired with a colleague at that institution. Thus, initially, each reviewer independently completed each of the 10 exercises; subsequently, the two reviewers—i.e., the "pair"—reconciled their independent responses through consensus or mediation by a third reviewer.

Reviewers were provided with the relevant key questions from the original CER; the treatment outcome being measured; the drug treatments being compared; and for each study, the study design, number of participants, study quality, and analysis results. One of the exercises included results from a meta-analysis. For each exercise, reviewers scored the four “required” strength of evidence domains (risk of bias, consistency, directness, and precision), separately for RCTs and observational studies. Then they used these domain-specific scores (and perhaps other information) to determine one grade for the overall strength of evidence for each outcome.

We also asked each reviewer to assess the difficulty of assigning a score for each domain in each exercise. They used a graduated scale of levels that ranged from “Very Easy” to “Very Difficult.”

For analytic purposes, we transcribed domain scores and overall strength of evidence grades into numeric responses and analyzed results quantitatively. We calculated two summary statistics to describe agreement among reviewers:

- The AC1 statistic (alternative chance-correlated coefficient, ranging from -1.00 to 1.00) measuring the agreement on the rating score among all reviewers across all exercises; and
- A statistic measuring agreement among reviewers that it was difficult or very difficult to determine the appropriate rating.

We used logistic regression analysis to assist in describing the relationship between domain scores and the strength of evidence grade, both in relation to the specific grade selected (i.e., insufficient; low; or moderate/high [combined as the comparison]) and level of agreement among reviewers in the grade selected. We examined the change in independent reviewer ratings following the reconciliation with the ratings of his or her partner.

To provide contextual insights into our quantitative analyses, we qualitatively synthesized the reasons that reviewers found particular rating exercises difficult.

Results

Overall, the level of agreement for domain scores was generally better across reviewers in their evaluations of RCTs than observational studies; exceptions were individuals’ ratings of precision and pairs of reviewers’ evaluations of directness. Likely because of the small sample sizes, virtually none of the differences in agreement reached statistical significance; the sole exception was the risk of bias domain. We did not discern a pattern in the correlation between agreement among individual reviewers and the percentage of reviewers who found the rating activity difficult or very difficult. Table A summarizes inter-rater testing results.

Table A. Inter-rater reliability results by domain and overall grade

Domain/ Strength of Evidence	Study Design	Number of Exercises	AC1 Agreement Across Independent Reviewers (95% CI)	Percentage of Independent Reviewers (SD) Describing Rating as Difficult or Very Difficult	Correlation Between Independent Reviewer Ratings on Agreement and Difficulty	Agreement Across Reconciled Pairs of Reviewers: AC1 (95% CI)
Risk of bias	RCT	10	0.67(substantial) (0.61 to 0.73)	3.2% (3.1%)	r=0.21	0.65 (substantial) (0.56 to 0.73)
	Observational	6	0.11 (slight) (0.05 to 0.18)	9.8% (6.0%)	r=-0.33	0.22 (fair) (0.13 to 0.32)
Consistency	RCT	10	0.51 (moderate) (0.34 to 0.67)	8.7% (9.0%)	r=-0.92	0.70 (substantial) (0.51 to 0.90)
	Observational	6	0.40 (fair) (0.13 to 0.66)	6.1% (8.0%)	r=-0.74	0.55 (moderate) (0.22 to 0.89)
Directness	RCT	10	0.73 (substantial) (0.60 to 0.87)	4.1% (4.0%)	r=0.09	0.78 (substantial) (0.64 to 0.92)
	Observational	6	0.48 (moderate) (0.32 to 0.64)	6.1% (8.5%)	r=-0.80	0.78 (substantial) (0.52 to 1.02)
Precision	RCT	10	0.23 (fair) (0.11 to 0.35)	17.8% (8.4%)	r=0.24	0.47 (moderate) (0.17 to 0.77)
	Observational	6	0.31 (fair) (0.06 to 0.56)	14.5% (3.4%)	r=0.09	0.38 (fair) (0.06 to 0.70)
Strength of Evidence	All exercises	10	0.20 (slight) (0.16 to.25)	19.6% (11.0%)	r=0.06	0.24 (fair) (0.14 to 0.34)
	RCTs only	4	0.22 (fair) (0.17 to 0.28)	14.7% (6.8%)	r=-0.09	0.30 (fair) (0.17 to 0.43)

Abbreviations: AC1 = alternative chance-correlated coefficient = an alternative to a kappa statistic; CI = confidence interval; r = correlation; RCT = randomized controlled trial; SD = standard deviation.

Key Question 1: Domain Scores Among Independent Reviewers

How consistent are domain score assessments conducted by individual independent reviewers (i.e., those done separately by a single reviewer)? Do inter-rater reliability calculations indicate patterns of reasonable agreement across reviewers?

The level of independent reviewer inter-rater agreement for domain scores varied considerably—from substantial for RCT risk of bias (AC1 = 0.67) and directness (AC1 = 0.73) to slight for observational study risk of bias (AC1 = 0.11). Agreement on all other domains was either moderate or fair.

Are any of the required domains more difficult or problematic for independent reviewers to assess than others?

For both RCTs and observational studies, agreement on precision was only fair. Reviewers expressed a desire for greater guidance when they could not rely on a meta-analysis and were faced with such problems as statistical significance expressed through p-values but not confidence intervals, a variety of differently measured outcomes, and nonsignificant findings. As presented in Table A, with the exception of the precision domain, only a small percentage of reviewers commented that they found rating specific domains to be “difficult” or “very difficult.”

Are domain scores for observational studies more difficult or problematic for independent reviewers to assess than those for RCTs?

Based on generally poorer inter-rater reliability results, we conclude that domain assessments for observational studies were more problematic than for RCTs. The particularly low agreement (slight) in relation to the risk of bias assessment of observational studies most likely relates to reviewers' not receiving sufficient guidance concerning the criteria the project CER teams had originally used for determining the quality of the studies and the different methodological approaches of the reviewers themselves.

Key Question 1: Domain Scores Among Reviewer Pairs

How consistent are domain scores that are the result of reconciliation by pairs of reviewers (i.e., assessment of scores on domains from two independent reviewers that are reconciled) across pairs of reviewers? Is the level of agreement among scores assessed by reconciled pairs greater than the level of agreement among domain scores assessed by independent reviewers?

Agreement on domain scores for reconciled reviewer pairs overall was as good as or better than it was for individual independent reviewers. Agreement on three of the four RCT scores was substantial (risk of bias, consistency, and directness). Agreement on appropriate precision domain scores was poorer than for the other domains, but it improved from fair to moderate for pairs. Agreement on domain scores for observational studies across reconciled pairs also improved in all domains except precision, but agreement was substantial only for directness. Based on these findings, we conclude that the reconciliation process is a critical step in domain scoring.

When reviewer pairs disagree on a domain score, in what direction do the reviewers generally reconcile their disagreements (e.g., toward better or toward worse domain scores)?

The direction of change in scores when a pair of reviewers had to settle a difference (i.e., had to reconcile their original scores) was inconsistent across domains and were reconciled to be "better" or "worse" in no obvious pattern.

Does the mechanism used by reviewer pairs to resolve disagreements in domain scores affect the agreed on score (i.e., does it matter whether the disagreement is resolved through consensus discussion between the two independent reviewers or through adjudication by a third reviewer)?

Only one pair of reviewers adjudicated differences by using a third party. Therefore, we had insufficient data to evaluate the effect of using consensus discussion versus a third party adjudicator on level of agreement across pairs.

Key Question 2: Strength of Evidence Grades Among Independent Reviewers

How consistent are strength of evidence grade assessments conducted by individual, independent reviewers (i.e., those done separately by a single reviewer)? Do inter-rater reliability calculations indicate patterns of reasonable agreement?

Agreement on independent reviewer strength of evidence grades overall was generally poorer than for domain scores. For the overall strength of evidence grades, inter-rater reliability agreement was slight among individual reviewers (AC1=0.20). Almost 20 percent of reviewers found the exercise of rating the strength of evidence to be difficult or very difficult.

Are particular domain scores more likely than others to be predictive of agreement in the overall strength of evidence grade?

We found that neither agreement on domain scores nor agreement about the level of difficulty that reviewers ascribed to evaluating particular domains predicted the overall grades.

Does agreement in strength of evidence grades differ by whether the evidence consists solely of RCTs or a combination of RCTs and observational studies?

Inter-rater reliability was not appreciably better when we considered only the four exercises limited to RCTs (AC1=0.22) as compared with all exercises (AC1=0.20).

When evidence was limited to RCT studies (four exercises), better strength of evidence grades of moderate/high (compared with both insufficient and low) were related to RCT domain scores' being considered consistent and precise. Because all RCT studies were presented as fair quality and were head-to-head trials, it is not surprising that the risk of bias and consistency domains were not predictors of the final strength of evidence grade.

Based on results from all 10 exercises, the inclusion of observational studies, in addition to RCTs, in an exercise was a strong predictor of a poorer strength of evidence grade —namely, either insufficient or low versus moderate/high.

Looking at the relationship between specific observational studies domain scores and strength of evidence grades, we found that observational study evidence being considered low risk of bias was significantly related to strength of evidence being graded as insufficient (but not low) versus moderate/high. This counterintuitive finding may be related to the findings in the two exercises with good-quality observational studies conflicting with the findings from the RCT evidence. By considering the observational studies as well as the RCT data, reviewers might reasonably have concluded that they could not reach a conclusion about the body of evidence. Also, consistency of observational studies data was positively related to the strength of evidence grade being moderate or high versus insufficient. This finding would seem to reflect reviewers' consideration of observational studies as secondary evidence that supports RCT findings when the direction of the findings is clear and not in conflict.

Does using different methods for combining domain scores into a single, overall strength of evidence grade result in a meaningful difference regarding the ultimate grade or in the time and effort expended? Possible methods include:

- The weighting system applied through using the GRADE algorithm (i.e., the approach promulgated by the GRADE Working Group),
- The EPC's own "numeric" or quantitative" weighting system, or
- The EPC's own "qualitative" approach to weighting or combining domain scores?

Methodological approaches to grading the strength of evidence were insufficiently described to be distinctly categorized and evaluated in relation to strength of evidence grading decisions.

Subgroup Analyses of Independent Reviewer Assessments of Domain Scores and Strength of Evidence Grades

Five subgroup analyses compared level of agreement by clinical condition (MDD versus RA), type of outcome (benefit versus harm), reviewers' experience in conducting systematic reviews, experience in evaluating strength of evidence, and academic training (physician versus nonphysician). Few differences were significant, and many were small.

Reviewer agreements by clinical condition and by type of outcome (benefits versus harms) were generally small and lacked informative patterns.

Reviewers with greater experience in conducting systematic reviewers (six of eight comparisons) and those with greater experience in evaluating strength of evidence (four of eight comparisons) were more likely to agree on domain scores. In contrast, reviewers with less experience (on both measures) were more likely to agree on strength of evidence grades than reviewers with greater experience.

We found few differences in agreement based on reviewers' type of academic training (i.e., physician or nonphysician). Because all participants were experienced reviewers from EPCs and because we did not evaluate background knowledge of the particular clinical conditions, we concluded that this distinction likely did not capture differences among reviewers in clinical and methodological expertise.

Key Question 2: Strength of Evidence Grades Among Reviewer Pairs

How consistent are strength of evidence grades that are the result of reconciliation by reviewer pairs (i.e., assessment of grades from two independent reviewers that are reconciled) across pairs of reviewers? Is the level of agreement among strength of evidence grades assessed by reconciled pairs greater than the level of agreement among strength of evidence grades assessed by independent reviewers?

Approximately 46 percent of strength of evidence grades were the same across independent reviewer pairs and so did not need to be reconciled.

Agreement on strength of evidence grades across reconciled pairs, compared with agreement for independent reviewers, improved modestly, from slight to fair, across exercises that had evidence from both RCTs and observational studies. It remained fair across exercises with only RCT evidence.

When reviewer pairs disagree on strength of evidence grades, in what direction do the reviewers generally reconcile their disagreements (e.g., toward better or toward worse strength of evidence grades)?

Final strength of evidence grades that needed to be reconciled were no more likely to be changed to a better (higher) or a worse (lower) grade. Approximately 25 percent were reconciled to a higher grade and approximately 30 percent to a lower grade. The pattern of reconciliation to better or worse final strength of evidence grades was not generally related to which of the two grades were being reconciled.

Does the mechanism used by reviewer pairs to resolve disagreements in strength of evidence grades affect the agreed-on grade (i.e., does it matter whether the disagreement is resolved through consensus discussion between the two independent reviewers or through adjudication by a third reviewer)?

We lacked sufficient data to determine whether the mechanism used to resolve disagreements between the two independent reviewers affected the final agreed-upon grade.

Discussion

Conclusions

The series of exercises we designed deliberately reflected the diversity and complexity of evidence that EPC reviewers encounter in their day-to-day evaluations. Our findings clearly demonstrate that the conclusions reached by experienced reviewers based on the same evidence can differ greatly. Consistency across reviewers can suffer when they are faced with complex bodies of evidence, especially when those data do not lend themselves to meta-analysis. Reflecting our analytic framework, we considered three factors that may have influenced the level of agreement that may have influenced the level of agreement on domain scores and final strength of evidence grades: reviewers' methodological approach, their judgment, and training.

Based on substantial agreement in domain scores and overall strength of evidence for the one exercise based on meta-analysis results from RCTs, current methodological guidance may be generally sufficient for straightforward evaluations that can rely on quantitative tools for summarizing the available information. In contrast, levels of agreement suffered when reviewers were faced with qualitative evaluations that did not lend themselves to meta-analysis and they needed to rely more heavily on their own judgment. Of particular concern is how to deal with (a) outcomes that are evaluated through a combination of RCTs and observational studies, (b) outcomes that are evaluated through more than one measure, and (c) grading evidence that appeared to show no difference. Reviewers need additional guidance on approaches to summarizing the strength of evidence, given various domain scores and combinations of such scores.

Additional training may also be desirable. In particular, we believe that some reviewers inappropriately considered some issues within incorrect domain categories. Training could provide relatively less experienced reviewers with greater knowledge of how to approach the various steps in the grading strength of evidence task. Over time, this enhanced educational effort should improve consistency.

Limitations

Our findings should be considered in light of several limitations: we limited the assessment to the four "required" domains in the AHRQ EPC guidance; we required reviewers to evaluate evidence about two clinical conditions for which they may have had limited or no prior knowledge; although we provided reviewers with p-values for exercises in which such information was missing, we did not calculate confidence intervals (when they were missing); and we did not give reviewers the criteria that the authors of the original CERs had applied to determine their own quality (or risk of bias) ratings of individual observational studies.

Needed Guidance Enhancements and Future Research

We conclude that additional methodological guidance is needed, including more details and examples, supported by more training, particularly on how best to evaluate the "thornier" bodies of evidence as discussed above. However, some potential will always exist for disagreement, even among experienced reviewers. For that reason, EPC reviewer teams need to be transparent in how they have conducted this task – that is, by documenting and describing their procedures. This will help to ensure that stakeholders can be confident of their understanding of a reviewer team's interpretation of the evidence and their ability to make decisions using such information.

Our study provided only a first approximation of reviewers' rationales for differences in domain scores and strength of evidence decisions. Additional research is needed to identify gaps in guidance that should be filled, areas of insufficient understanding of the guidance itself and how best to overcome that deficit, and complex decisions that may still need to be left to the review team's substantive expertise. A future reliability study could compare whether a single, more standardized approach to grading strength of evidence, particularly arriving at an overall grade from domain scores, would provide greater reliability than the varied approaches that EPCs have been permitted to use in the current guidance. If inter-rater reliability is similar when methods are constrained to stipulated approaches that are less discretionary, then such findings might indicate not only that gaps remain in both approaches but also that, for complex evaluations, no "right" approach may exist. That, in turn, highlights the importance of transparency and the need for adequate explanation that speaks to the needs of all stakeholders.

Introduction

Background

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Center (EPC) program, sponsors Comparative Effectiveness Reviews (CERs) and other systematic reviews (SRs) of the literature. To advance this effort, AHRQ developed methodological guidance and supported related research to enhance the scientific rigor of these reviews.[1] AHRQ's guidance encompasses many of the required tasks of a review.

Chapters of the Methods Guide for Effectiveness and Comparative Effectiveness Reviews (Methods Guide) cover a wide array of topics. Among them are developing analytic frameworks to describe the relationship between treatment options and health outcomes; searching for relevant evidence; using observational studies in addition to randomized trials; evaluating the risk of bias (quality) of individual included studies; quantitatively synthesizing evidence; grading the strength of the evidence in relation to outcomes; and assessing the applicability of findings from the included studies.

The guidance on the methodological approach to employ in completing these various tasks is developed through working groups, comprised primarily of expert practitioners from the independent EPCs. One such cross-EPC team developed the chapter on grading the strength of evidence related to therapeutic interventions, which was published in 2010 as definitive guidance for EPCs.[2]

We report here on an evaluation of the inter-rater reliability of the strength of evidence guidance when used by experienced reviewers across the EPCs. We conducted this research starting in 2010 (after the publications above were available). Our primary goal was to determine whether different teams of reviewers would reach similar conclusions on the strength of evidence when presented with the same information about studies included in a CER. Our second goal was to gain a greater understanding of the relative role of each of the criteria (referred to as domains) that are evaluated in developing a strength of evidence grade.

AHRQ EPC Approach to Evaluating Strength of Evidence

The strength of evidence grade is a key indicator of a review team's level of confidence that the studies included in the review collectively reflect the true effect of an intervention on a health outcome. The AHRQ Methods Guide chapter for grading the strength of evidence instructs reviewers to score the body of evidence in relation to each major outcome (e.g., benefits and harms) and each main comparison (e.g., intervention A vs. intervention B) in relation to each key question of the review.[2]

First, two independent reviewers evaluate (score) critical domains. The four required domains include the risk of bias of included studies, the consistency of the evidence, the directness of the evidence, and the precision of the estimates (Table 1). Four additional domains that can be scored if reviewers consider them to be integral to making a final evaluation include dose-response association, plausible confounding that would decrease observed effect, strength of association/magnitude of effect, and publication bias. The Methods Guide recommends scoring each of the domains separately for randomized controlled trials (RCTs) and observational (nonrandomized) studies that are evaluated for each outcome. Disagreements between the two reviewers in their decisions concerning domains scores are resolved through consensus or adjudication by a third reviewer.

Table 1. Required domains in the AHRQ EPC approach to grading strength of evidence*

		Score
		Bolded score: raises strength of evidence
Domain	**Key Elements**	***Italicized* score: lowers strength of evidence**
Risk of bias	Degree to which included studies for specified outcome & comparison have a high likelihood of adequate protection against bias (i.e., good internal validity), assessed through two main elements: study design and aggregate quality of included studies being considered (based on quality or risk of bias rating of individual studies [good/fair/poor]).	Scored as one of three levels: **Low** Medium *High* If included studies differ in risk of bias, greater weight can be given to those with lower risk
Consistency	Degree to which reported effect sizes of included studies for specified outcome and comparison have same direction of effect. Assessed by two main elements: effect sizes have same sign (i.e., are on same side of no effect) and range of effect sizes is narrow	Scored as one of three levels: **Consistent:** no inconsistency *Inconsistent:* non-overlapping confidence intervals, significant unexplained clinical or statistical heterogeneity, statistically significant effect sizes in opposite directions *Unknown or not applicable*: includes single-study evidence
Directness	Whether the evidence links the intervention directly to a health outcome. If two treatments are compared, directness implies that head-to-head trials measure the most important outcome. Evidence can be indirect in two ways: (1) Uses intermediate or surrogate outcomes rather than ultimate health outcomes; (2) Uses two or more bodies of evidence to compare interventions A and B without having head-to-head studies of A vs. B	Scored as one of two levels: **Direct** *Indirect*: specify which of the two types, or both if applicable. Note potential weaknesses caused by indirect evidence or analysis
Precision	Degree of certainty surrounding an effect estimate for a specified outcome and comparison. If meta-analysis is used, it is the confidence interval around the summary effect size.	Scored as one of two levels: **Precise**: clinically useful conclusion *Imprecise:* would not allow for clinically distinct conclusions such as clinical inferiority

* From Owens, et al., 2010.[2]

Finally, for each major outcome and each major comparison (e.g., intervention A vs. intervention B), EPCs then aggregate domains scores into a single strength of evidence grade (Table 2). The Methods Guide does not dictate one way for EPCs to incorporate the multiple domain scores to arrive at the overall strength of evidence grade in relation to an outcome. EPCs can use the algorithm developed by the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group,[3] their own weighting (e.g., numeric) system, or a qualitative approach. The only requirement is transparency; EPCs must clearly explain their rationale for aggregating the domains in a single strength of evidence grade.

Table 2. Strength of evidence grades and definitions

Grade	Definition
High	**High confidence that the evidence reflects the true effect.** Further research is very unlikely to change our confidence in the estimate of effect.
Moderate	**Moderate confidence that the evidence reflects the true effect.** Further research may change our confidence in the estimate of effect and may change the estimate.
Low	**Low confidence that the evidence reflects the true effect.** Further research is likely to change our confidence in the estimate of effect and is likely to change the estimate.
Insufficient	Evidence either is unavailable or does not permit estimation of an effect.

* From Owens, et al., 2010.[2]

The authors of the Methods Guide chapter on strength of evidence concluded that refinements to the current guidance would likely warrant cross-EPC attention and additional empirical work in the years ahead. Thus, the present guidance is understood to be open for modification as more experience with research into these issues accumulates.

Background: Quality of Evidence Reliability Project Conducted by the GRADE Working Group

Several years ago, the GRADE Working Group tested the inter-rater reliability of a pilot version of their approach to grading strength of evidence (referred to by GRADE as quality of evidence) and recommendations.[3] In that study, researchers provided an evidence profile to reviewers, consisting of two tables. One table presented a quality assessment of the available information for each outcome and included the following: the number of studies that were included to evaluate the outcome, the type of study design (RCT or observational), the quality of each included study, consistency of results across studies, and directness of the available evidence. The second table presented a summary of findings for each outcome: data on the number of patients and events and measures of the relative effect of treatments.

Seventeen reviewers evaluated 12 outcomes (i.e., specific research questions) drawn from a single systematic review. The reviewers graded the strength of evidence for each evidence profile based solely on the information in the evidence profiles. Kappa agreement among raters for the 12 outcomes was fair (κ=0.27 [standard error (SE), 0.015]), ranging from agreement slightly worse than chance for four outcomes (negative kappa values) to a high of κ=0.823 for one outcome.[4,5]

Study Objectives

Our project focused on inter-rater reliability testing of the two main components of the AHRQ approach to the task of grading strength of evidence for specific outcomes in relation to key questions: (1) scoring evidence on the four required domains (risk of bias, consistency, directness, and precision) and (2) developing an overall strength of evidence grade, given the scores for these four individual domains.

Thus, our research focused on key questions concerning the performance of these two main tasks:

Key Question 1. Domain Scores

Independent Reviewers

- How consistent are domain score assessments conducted by individual independent reviewers (i.e., those done separately by a single reviewer)? Do inter-rater reliability calculations indicate patterns of reasonable agreement across reviewers?
- Are any of the required domains more difficult or problematic for independent reviewers to assess than others?
- Are domain scores for observational studies more difficult or problematic for independent reviewers to assess than those for RCTs?

Reviewer Pairs

- How consistent are domain scores that are the result of reconciliation by pairs of reviewers (i.e., assessment of scores on domains from two independent reviewers that are reconciled) across pairs of reviewers? Is the level of agreement among scores assessed by reconciled pairs greater than the level of agreement among domain scores assessed by independent reviewers?
- When reviewer pairs disagree on a domain score, in what direction do the reviewers generally reconcile their disagreements (e.g., toward better or toward worse domain scores)?
- Does the mechanism used by reviewer pairs to resolve disagreements in domain scores affect the agreed on score (i.e., does it matter whether the disagreement is resolved through consensus discussion between the two independent reviewers or through adjudication by a third reviewer)?

Key Question 2. Overall Strength of Evidence Grade

Independent Reviewers

- How consistent are strength of evidence grade assessments conducted by individual, independent reviewers (i.e., those done separately by a single reviewer)? Do inter-rater reliability calculations indicate patterns of reasonable agreement?
- Are particular domain scores more likely than others to be predictive of agreement in the overall strength of evidence grade?
- Does agreement in strength of evidence grades differ by whether the evidence consists solely of RCTs or a combination of RCTs and observational studies?
- Does using different methods for combining domain scores into a single, overall strength of evidence grade result in a meaningful difference regarding the ultimate grade or in the time and effort expended? Possible methods include:
  - The weighting system applied through using the GRADE algorithm (i.e., the approach promulgated by the GRADE Working Group),
  - The EPC's own "numeric" or quantitative" weighting system, or
  - The EPC's own "qualitative" approach to weighting or combining domain scores?

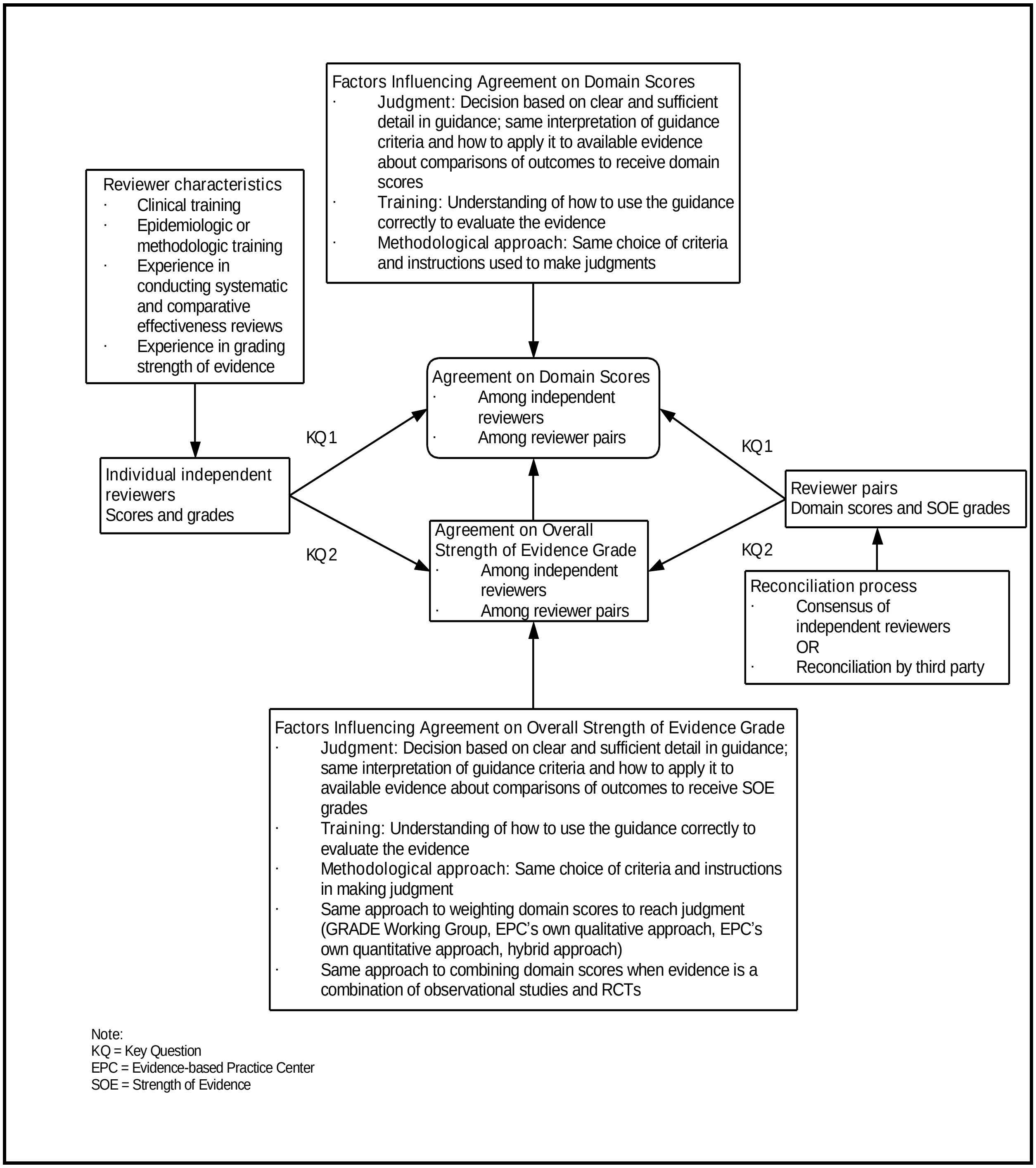
Reviewer Pairs

- How consistent are strength of evidence grades that are the result of reconciliation by reviewer pairs (i.e., assessment of grades from two independent reviewers that are reconciled) across pairs of reviewers? Is the level of agreement among strength of evidence grades assessed by reconciled pairs greater than the level of agreement among strength of evidence grades assessed by independent reviewers?

- When reviewer pairs disagree on strength of evidence grades, in what direction do the reviewers generally reconcile their disagreements (e.g., toward better or toward worse strength of evidence grades)?
- Does the mechanism used by reviewer pairs to resolve disagreements in strength of evidence grades affect the agreed-on grade (i.e., does it matter whether the disagreement is resolved through consensus discussion between the two independent reviewers or through adjudication by a third reviewer)?

Our analytic framework (Figure 1) displays linkages between our population (individual independent reviewers and reviewer pairs) and outcomes (agreement on domain scores and overall strength of evidence grades). We present reviewers' characteristics (background and experience) that may affect their independent scores and resulting grades; we also note reviewer pairs' reconciliation approach because it may affect their reconciled scores and grades. We also present factors that may influence level of agreement across independent reviewers and reviewer pairs (differences in reviewers' judgment, training, and methodological approach). The discussion chapter is organized to synthesize our conclusions in relation to our Key Questions and comment on the relative influence of these three factors.

Figure 1. Analytic framework for inter-rater reliability in EPC approach to grading strength of evidence


Factors Influencing Agreement on Domain Scores
· Judgment: Decision based on clear and sufficient detail in guidance; same interpretation of guidance criteria and how to apply it to available evidence about comparisons of outcomes to receive domain scores
· Training: Understanding of how to use the guidance correctly to evaluate the evidence
· Methodological approach: Same choice of criteria and instructions used to make judgments
Reviewer characteristics
· Clinical training
· Epidemiologic or methodologic training
· Experience in conducting systematic and comparative effectiveness reviews
· Experience in grading strength of evidence
Agreement on Domain Scores
· Among independent reviewers
· Among reviewer pairs
KQ1
KQ1
Individual independent reviewers
Scores and grades
Reviewer pairs
Domain scores and SOE grades
KQ2
KQ2
Agreement on Overall Strength of Evidence Grade
· Among independent reviewers
· Among reviewer pairs
Reconciliation process
· Consensus of independent reviewers
OR
· Reconciliation by third party
Factors Influencing Agreement on Overall Strength of Evidence Grade
· Judgment: Decision based on clear and sufficient detail in guidance; same interpretation of guidance criteria and how to apply it to available evidence about comparisons of outcomes to receive SOE grades
· Training: Understanding of how to use the guidance correctly to evaluate the evidence
· Methodological approach: Same choice of criteria and instructions in making judgment
· Same approach to weighting domain scores to reach judgment (GRADE Working Group, EPC's own qualitative approach, EPC's own quantitative approach, hybrid approach)
· Same approach to combining domain scores when evidence is a combination of observational studies and RCTs

Note:
KQ = Key Question
EPC = Evidence-based Practice Center
SOE = Strength of Evidence

Methods

Information From Recent Comparative Effectiveness Reviews

We conducted the inter-rater reliability testing using data from two published CERs focusing on two distinct chronic conditions and drug treatments: second-generation antidepressants for the treatment of major depressive disorder (MDD) and disease-modifying antirheumatic drugs (DMARDs) for the treatment of rheumatoid arthritis (RA).[6,7] From the data in these reviews, the study team designed 10 exercises, which are summarized in Table 3. All 10 exercises included RCTs, and 6 of the 10 also included one or more observational studies. Using the same terminology as is used in the AHRQ Methods Guide chapter on strength of evidence grading, observational studies include nonrandomized trials, cohort, cross-sectional, and case-control studies. The risk of bias of individual studies was provided as a quality score (good, fair, or poor) because this was the metric used by the original reviewers. These exercises together evaluated a variety of efficacy (benefit) and adverse event (harm) outcomes across the two conditions.

The exercises also provided a range of intellectual challenges relating to evaluating different aspects of the strength of evidence guidance (see right column of Table 3). These challenges included:

- data that did not lend themselves to meta-analysis; only one exercise (response on the Hamilton Rating Scale for Depression [HAM-D]) evaluated pooled data through results from a meta-analysis of RCTs;
- outcomes evaluated with different measures across studies;
- inconsistent lengths of time across studies for when outcomes were measured;
- outcomes evaluated through both RCTs and observational studies;
- limited information in studies, such as the inclusion of p-values but no confidence intervals;
- differences in sample sizes across studies; and
- differences in study quality (risk of bias).

Table 3. Grading exercises: description and main characteristics of evidence

Condition Outcome	RCTs Number of Studies[a] (Total N across studies)	Observational Studies Number of Studies (Total N across studies)	Main Characteristics and Challenges of the Exercise
MDD Benefits: • HAM-D response	5 studies[8-12] (N=690)	None	All RCTs with similar results Evaluation of pooled data (meta-analysis results) Exercise considered straightforward by study team
MDD Benefits: • Efficacy response in subpopulation of elderly	1 study[13] (N=108)	None	Limited to one RCT evaluated through 2 outcome measures Results in bar graph only (no exact scores) Precision presented through p-values only (no confidence intervals) Exercise considered straightforward by study team

Table 3. Grading exercises: description and main characteristics of evidence (continued)

Condition Outcome	RCTs Number of Studies[a] (Total N across studies)	Observational Studies Number of Studies (Total N across studies)	Main Characteristics and Challenges of the Exercise
MDD Harms: - Sexual dysfunction	4 studies[8,10,11,14] (N=904	2 studies[15,16] (1 prospective cohort, 1 large cross-sectional survey) (N=3,154)	Mix of study designs: 4 RCTs (including 1 open-label), 2 observational studies Outcome measure differed across studies Indirect measure in cross-sectional study Exercise considered challenging by study team
MDD Harms: - Suicidality	1 study[12] (N=90)	2 studies[17,18] (1 case-control and 1 large good-quality nested case-control) (N=11,350)	Mix of study designs: 1 RCT, 2 observational studies Only RCT found difference Outcome measures differed across studies, some measure ideation Exercise considered challenging by study team
MDD Harms: - Nausea	5 studies[8,9,11-13] (N=689)	None	All RCTs No significant differences between arms in any study. In 4 studies, paroxetine has higher rates and in 1 study fluoxetine has a higher rate. Exercise considered straightforward by study team
RA Benefits: - ACR20 response	5 studies[19-27] (N=1,639)	2 studies[28,29] (1 nonrandomized trial, 1 prospective cohort) (N=2,461)	Mix of study designs: 5 RCTs, 2 observational studies Different followup durations Mixed results Composite outcome in 1 study Exercise considered challenging by study team
RA Benefits: - ACR70 response	5 studies[19-27] (N=1,639)	1 study[28] (nonrandomized open-label trial) (N=269)	5 RCTs, 1 observational study Only 1 study found a difference between treatments Different followup durations Same included studies as ACR20 exercise plus the addition of an observational study Exercise considered challenging by the study team
RA Benefits: - DAS remission	2 studies[20,23,25-27] (N=982)	1 study[30] (prospective cohort) (N=1,083)	Mix of study designs; 2 RCTs did not find a difference, but 1 large observational study did. Exercise considered challenging by study team
RA Harms: - Serious infection	4 studies[20-23,25-27] (N=1,215)	2 studies[31,32] (1 large fair retrospective cohort and 1 good retrospective cohort) (N=7,695)	Mix of study designs: 4 fair RCTs, 2 observational studies Difference found in only 1 large retrospective study Outcome is rare, so few events Exercise requires incorporating retrospective data Exercise considered challenging by study team
RA Harms: - Infusion or injection reaction	4 studies[19,21-27] (N=1,108)	None	All RCTs Differences seen in 3 of 4 studies; no difference found in the smallest study Exercise considered straightforward by study team

[a] Numbers of citations may exceed number of studies when the latter are published in multiple articles. Abbreviations: ACR = American College of Rheumatology; CI = confidence interval; DAS = Disease Activity Scale; HAM-D = Hamilton Rating Scale for Depression; MDD = major depressive disorder; N = number; RA = rheumatoid arthritis; RCT = randomized controlled trial.

The project was not a perfect replica of the actual data confronted and processes undertaken by EPC investigators in grading the strength of a body of evidence in a CER or systematic review. Instead, we attempted to replicate the key parts of the process as closely as possible, acknowledging time constraints of the reviewers. Our goal was to craft a dataset that addressed a wide range of factors in grading strength of evidence but that did not require either prior knowledge of the subject matter or an unreasonable amount of time to complete the exercises in a thoughtful and meaningful way. Thus, we did not replicate the analysis conducted through the original CERs; this project should not be considered a reevaluation of the results presented in the original reports.

For MDD, we focused solely on treatment comparisons between two specific second-generation antidepressants: fluoxetine and paroxetine. We selected two efficacy and three adverse event outcomes (see Table 3 above). For RA, because of the limited number of eligible head-to-head studies comparing the same individual drugs, we selected a set of studies comparing any biologics with any oral DMARDs. We instructed reviewers to treat the biologics as equivalent and the oral DMARDs as equivalent and to ignore any potential differences between drugs within a class. For RA, we selected three efficacy and two adverse event outcomes.

Information Provided to Reviewer Participants

In October 2010, we provided reviewers with a summary table for each of the 10 exercises. This table contained information that we considered essential for completing the exercises. Specifically it provided: the relevant key question from the original CER; the treatment outcome being measured; the drug treatments being compared; and, for each study, the study design, number of participants, study quality (good or fair—none of the studies included in the exercise was rated as poor quality), and analysis results (Appendix A). We provided reviewers with other essential materials, including detailed instructions on how to complete the exercises (Appendix B), the chapter of the Methods Guide on grading the strength of a body of evidence,[2] and background documents to provide baseline knowledge of MDD and RA, including descriptions of the disease conditions, treatment mechanisms and outcomes measures that would be examined in the exercises. We also sent reviewers supplemental materials; these included evidence tables and full-text articles for each study in the exercises. We instructed participants to use the supplemental materials if they wished; however, we designed the exercises so that they could be completed based on the summary tables and other essential materials and strength of evidence guidance document alone.

Participants

We invited all EPC Directors and our AHRQ Task Order Officer to nominate pairs of reviewers who were each experienced in evaluating strength of evidence. They did not need to have background knowledge of MDD or RA and could not have participated in the original CERs from which the data were obtained. A convenience sample consisting of 22 participants (11 pairs of reviewers) participated in the exercises. Two reviewers were from AHRQ; 20 came from nine EPCs, with eight EPCs each having two reviewers and one EPC having four, (see acknowledgments). To replicate "real world" or "real EPC" practices, we organized the commonly used approach for "dual assessment." Initially, each reviewer at a given institution independently completed each of the exercises; subsequently, the two reviewers—i.e., a "pair"—

reconciled their independent responses. (For the one EPC with four reviewers, the EPC established the "pairs.")

Data Collection

We conducted the data collection effort during October and November 2010. The project team created electronic response forms to collect information from reviewers through ZipSurvey™, an online survey program. We used the electronic response forms to collect three types of information from the 22 reviewers: (1) background information on reviewers--their experience with conducting systematic reviews and grading strength of evidence and whether they had been trained as a physician; (2) domain scoring and strength of evidence grading results; and (3) qualitative feedback on conducting the assessment. The 11 pairs of reviewers also used the electronic response forms to submit their reconciled evaluations and any feedback to us.

For each of 10 exercises, we asked reviewers to complete two steps:

1. Score the four "required" strength of evidence domains (risk of bias, consistency, directness, and precision), separately for RCTs and observational studies, and then
2. Use this information to develop one overall strength of evidence grade for each outcome.

Reviewers were also asked to assess the difficulty of assigning a score for each domain in each exercise using a five level graduated scale that ranged from "Very Easy" to "Very Difficult." If reviewers assessed a domain score as being either "Difficult" or "Very Difficult," they were directed to a text box to elaborate on their response.

We also asked reviewers to provide a summary of their results for each exercise; this included noting whether one of the treatment arms was considered superior and important similarities or differences among studies. As above, we asked them to distinguish findings for RCTs and observational studies, if applicable.

Finally, we solicited reviewers' feedback on their overall experience in assigning domain scores and a strength of evidence grade for each exercise. Reviewer pairs were asked to state the process of reconciliation that they used, to comment on any domains or outcomes that they found especially difficult to reconcile, and to provide any other feedback that they thought would be useful.

Data Synthesis and Analysis

We calculated two summary statistics to describe agreement among reviewers on each of the four domains and the overall strength of evidence grade. Because reviewers were given a predetermined set of choices, for analytic purposes we were able to transcribe domain scores and overall strength of evidence grades into numeric responses and analyze results quantitatively.

We computed the summary statistics for each domain separately across RCT studies (10 exercises) and observational studies (6 exercises). Each summary statistic for the strength of evidence grades was based on all studies in all 10 exercises.

Agreement on Rating Across Reviewers

The first summary measure is the first-order agreement coefficients (AC1 statistic, for alternative chance-correlated coefficient),[33] which measures the agreement on the rating score among all reviewers across all relevant exercises. The AC1 statistic is a summary measure for inter-rater reliability tests with multiple raters; it is similar to the commonly used kappa statistic.[34]

We selected the AC1 statistic over the kappa statistic because of the concerns about the so-called "kappa paradox," where high agreement can accompany low kappa scores.[35] AC1 overcomes this paradox by adjusting for chance agreement; it is considered an alternative and appropriate measure of inter-rater reliability.

The AC1 score ranges from -1.00 (no agreement) to 1.00 (100 percent agreement). We also calculated the 95% confidence interval (CI) corresponding to the estimate.

To assist in describing and interpreting the AC1 statistic, we adopted the scale developed by Landis and Koch[4] for interpreting the kappa statistic.[4,5] We describe six levels of AC1 agreement as follows: <0 (less than chance agreement); 0.01–0.20 (slight agreement); 0.21–0.40 (fair agreement); 0.41–0.60 (moderate agreement); 0.61–0.80 (substantial agreement); and 0.81–0.99 (almost perfect agreement).

Agreement on Difficulty Across Reviewers

The second summary statistic measures agreement among reviewers that it was "difficult" or "very difficult" to determine the appropriate domain score (separately for RCTs and observational studies) or the overall strength of evidence grade. We computed this summary statistic in two steps. For each domain (separately by type of study—RCT or observational) and for overall strength of evidence, we first computed the percentage of reviewer ratings of difficult or very difficult for each of the exercises. We averaged these percentages across all exercises to produce a summary statistic, and we then calculated the standard deviation of the summary statistic across the exercises. In addition, we computed the correlation coefficient between the percentage of reviewers who agreed on each domain score and the percentage of reviewers who considered the rating exercise to be difficult or very difficult.

Prediction of Strength of Evidence Grade

We used logistic regression analysis to assess the relationship between domain scores and the strength of evidence grade at the individual-reviewer level. For these analyses, we used the SUDAAN MULTILOG procedure with the generalized estimating equation option (GEE) to control for repeated measures from the same reviewer. We estimated four models that were determined a priori. They differed in whether the dependent variable was an actual strength of evidence grade or a level of agreement on the strength of evidence grade and, secondarily, whether the data set comprised only six or all 10 exercises. Specifically they were:

- Model 1, a multinomial logistic regression model predicting strength of evidence grade by domain scores. The model included results from just the six exercises that included both RCTs and observational studies. The outcome (dependent variable) was the strength of evidence grades: insufficient; low; or moderate/high (combined as the comparison). This model had 131 observations (22 raters times 6 exercises, with 1 observation missing) and 12 independent variables, namely each of the domain scores, separately for RCTs and observational studies.
- Model 2, a modification of the above multinomial logistic regression model predicting strength of evidence grade by domain scores. This model included results from all 10

exercises by adding a dummy variable indicating whether an exercise had included observational studies, in addition to 6 independent variables for each of the RCT domain score ratings. It included 219 observations (22 raters times 10 exercises; 1 observation missing).

- Model 3, a logistic regression model predicting agreement on strength of evidence grade by agreement on domain scores. Agree was calculated as a dichotomous variable that was coded as 1 when the reviewer's response was in agreement with the most popular response among all reviewers; otherwise it was coded as 0. The model (like Model 1) includes results from just the six exercises that included RCTs and observational studies: 131 observations (1 observation missing) and 8 independent variables. The outcome (dependent variable) was agree; disagree (comparison) on the strength of evidence grade. The independent variables were agree; disagree (comparison) on each of the domain scores, separately for RCTs and observational studies.
- Model 4, a modification of the above logistic regression model predicting agreement on strength of evidence grade by agreement on domain scores. This model (like Model 2) included results from all 10 exercises by adding a dummy variable indicating whether an exercise included observational studies. It included 219 observations and 5 independent variables.

Subgroup Analyses of Independent Reviewer Assessments

We conducted subgroup analyses to examine whether reviewers' responses differed across domain scores and strength of evidence grades by several variables: clinical condition (MDD vs. RA); type of outcome (benefit vs. harm); academic training of the reviewer (physician vs. nonphysician); years of experience in conducting systematic reviews; and years of experience in grading strength of evidence. We combined all the reviewers' ratings across all exercises and computed the average agreement, stratified by subgroup. We tested whether the likelihood of a reviewer's response agreed with the most popular response, through logistic regression analysis using the SUDAAN MULTILOG procedure with GEE to control for repeated measures from the same reviewer.

In each subgroup analysis, we specified the reviewer's response as the dependent variable (coded as 1 if the response was the most popular response and 0 otherwise) and the stratification (subgroup) of interest (i.e., clinical topic, outcome type, academic training, and two measures of experience) as the one independent variable. For example, the analysis examining the effect of being a physician on the agreement for each domain score was specified as follows:

$\boldsymbol{logit\ (uij) = \alpha + \beta\ MD\_indicator_{ij}}$ i = 1 to 22 raters, j = 1 to 10 for RCT studies and 1 to 6 for observational studies; where

is 1 if the rater's response for the domain agreed with the most popular response and 0 otherwise,

$\boldsymbol{MD\_indicator_{ij}}$ is 1 if the rater i for the study j is a physician and 0 otherwise,

$\boldsymbol{\alpha}$ is the intercept, and

$\boldsymbol{\beta}$ is the slope of each logit model.

In addition to examining the significant results from each of the logistic regression analyses, we also examined patterns of percentage agreement (regardless of statistical significance) because of concerns about limitations in power resulting from a small sample size.

Reconciliation Across Reviewer Pairs

We examined changes in each independent reviewer's domain scores and in strength of evidence grades following reconciliation with his or her partner. We assigned a numeric value to each of the domains/strength of evidence responses for each reviewer and for each pair. For the consistency domain, we combined scores of inconsistent and unknown for this analysis because they were considered worse scores than consistent, but the two could not be clearly ranked with each other. A higher value indicated a better score and a lower value a worse score as follows:

- Risk of bias: Low = 3, Medium = 2, High = 1;
- Consistency: Consistent = 2, Inconsistent or Unknown = 1;
- Directness: Direct = 2, Indirect = 1;
- Precision: Precise = 2; Imprecise = 1;
- Strength of evidence: High = 4, Moderate = 3, Low = 2, Insufficient = 1.

For each domain, we then subtracted each individual's score from the reconciled ('pairs') score, yielding two "difference scores." Next, we summed these two difference scores for each pair and classified the summed value into three change groups: agree (i.e., no change), better score, or worse score. Results for domain scores for RCT studies and strength of evidence were based on 11 pairs evaluating 10 outcomes (N=110). Results for domain scores for observational studies were based on 11 pairs evaluating 6 outcomes (N=66).

The majority of the change groups with a value of zero constituted the case in which the numeric value of the reconciled score and both of the individual scores was equal, indicating that the two individuals agreed on the domain response. As illustrated in Table 4 (example 3), both reviewers scored risk of bias as 2, so subtracting each of their scores from the reconciled score, which by definition was 2, yielded a difference of zero. We put these into the "agree" group. In a few cases, however, the reconciled score was the mid-point of the two individual scores (example 4). For example, a reconciled score had a value 2 whereas the two individual scores were 1 and 3. Because this case has no clear direction of change, we grouped it also with the "agree" group.

Finally (see Table 4), a change score is positive (better) when the reconciled score was a larger value than either one or both of the individual scores; that is, this case indicated that reconciled score changed toward a "better" score direction. Similarly, a negative change score indicated that the move was toward a "worse" direction.

Table 4. Illustration of calculation of change scores

Example	Reconciled Score	Score From Reviewer #1	Score From Reviewer #2	Difference (Change) Score	Direction of Change
1	3	1	2	(3-1) + (3-2) = 3	Better score
2	2	3	2	(2-3) + (2-2) = -1	Worse score
3	2	2	2	(2-2) + (2-2) = 0	Agree/No change
4	2	1	3	(2-1) + (2-3) = 0	Agree/No change

Qualitative Analyses

We qualitatively synthesized the reasons that reviewers found particular domain scoring and strength of evidence grading difficult and present these findings in the results sections containing corresponding data from the inter-rater reliability testing. We also documented the approaches that reviewer pairs used to resolve disagreements. This information is intended to provide contextual insights into the results across reviewers.

Results

Inter-rater reliability testing results are presented first for each of the four domains. We specify the results separately for RCTs and observational studies. These findings are followed by results for the overall strength of evidence.

As shown in Table 5, across independent reviewers and reconciled pairs, the level of agreement (AC1) for domain scores was generally better in their evaluations of RCTs than observational studies. Likely because of the small sample sizes, only one of the comparisons of differences in level of agreement between RCTs and observational studies reached statistical significance. This occurred for risk of bias domain evaluations by both individual reviewers and pairs, as shown by nonoverlapping confidence intervals.

Table 5. Inter-rater reliability results by domain and overall grade

Domain/ Strength of Evidence	Study Design	Number of Exercises	AC1 (95% CI)	Percentage (SD) Describing Rating as Difficult or Very Difficult	Correlation Between Ratings on Agreement and Difficulty	Agreement Across Reconciled Pairs of Reviewers: AC1 (95% CI)
Risk of bias	RCT	10	0.67(substantial) (0.61 to 0.73)	3.2% (3.1%)	r=0.21	0.65 (substantial) (0.56 to 0.73)
	Observational	6	0.11 (slight) (0.05 to 0.18)	9.8% (6.0%)	r=-0.33	0.22 (fair) (0.13 to 0.32)
Consistency	RCT	10	0.51 (moderate) (0.34 to 0.67)	8.7% (9.0%)	r=-0.92	0.70 (substantial) (0.51 to 0.90)
	Observational	6	0.40 (fair) (0.13 to 0.66)	6.1% (8.0%)	r=-0.74	0.55 (moderate) (0.22 to 0.89)
Directness	RCT	10	0.73 (substantial) (0.60 to 0.87)	4.1% (4.0%)	r=0.09	0.78 (substantial) (0.64 to 0.92)
	Observational	6	0.48 (moderate) (0.32 to 0.64)	6.1% (8.5%)	r=-0.80	0.78 (substantial) (0.52 to 1.02)
Precision	RCT	10	0.23 (fair) (0.11 to 0.35)	17.8% (8.4%)	r=0.24	0.47 (moderate) (0.17 to 0.77)
	Observational	6	0.31 (fair) (0.06 to 0.56)	14.5% (3.4%)	r=0.09	0.38 (fair) (0.06 to 0.70)
Strength of Evidence	All exercises	10	0.20 (slight) (0.16 to.25)	19.6% (11.0%)	r=0.06	0.24 (fair) (0.14 to 0.34)
	RCTs only	4	0.22 (fair) (0.17 to 0.28)	14.7% (6.8%)	r=-0.09	0.30 (fair) (0.17 to 0.43)

Abbreviations: AC1 = alternative chance-correlated coefficient = an alternative to a kappa statistic; CI = confidence interval; r = correlation; RCT = randomized controlled trial; SD = standard deviation.

Among individual reviewers and reviewer pairs, the level of agreement on the strength of evidence grade was fair for evaluations that solely included RCTs. By contrast, agreement was slight among individual reviewers and fair among reviewer pairs for evaluations that included both RCTs and observational studies.

We did not discern any patterns in the correlation between agreement among individual reviewers and the percentage of reviewers who found the rating difficult or very difficult. Agreement was greater among reconciled pairs of reviewers than it was among individual reviewers; the sole exception (very small) was for risk of bias for RCTs. We present greater detail on results specific to each of the domains and strength of evidence below.

Risk of Bias

We had provided the reviewers with an assessment (rating) of the quality of each included study; for the two CERs used for this work, the authors had rated "quality" (rather than "risk of bias") using longstanding formal practices. All RCTs included in the 10 exercises had been rated fair quality (Table 6).

Table 6. Summary of RCT risk of bias domain scores, by condition and outcome

Condition Outcome	Number of Studies (Total N of subjects)	Study Quality Ratings	Individual Domain Scores	Reconciled Domain Scores
MDD Benefits: • HAM-D response	5 (N=690)	All Fair	Low: 14% Medium: 86% High: 0%	Low: 18% Medium: 82% High: 0%
MDD Benefits: • HAM-D & MADRS response in elderly	1 (N=108)	Fair	Low: 5% Medium: 73% High: 23%	Low: 0% Medium: 73% High: 27%
MDD Harms: • Sexual dysfunction	4 (1 open label) (N=904)	All Fair	Low: 14% Medium: 82% High: 4%	Low: 9% Medium: 91% High: 0%
MDD Harms: • Suicidality	1 (N=90)	Fair	Low: 0% Medium: 82% High: 18%	Low: 0% Medium: 64% High: 36%
MDD Harms: • Nausea	5 (N=689)	All fair	Low: 14% Medium: 86% High: 0%	Low: 18% Medium: 82% High: 0%
RA Benefit: • ACR20	5 (N=1,639)	All fair	Low: 14% Medium: 82% High: 4%	Low: 9% Medium: 91% High: 0%
RA Benefit: • ACR70	5 (N=1,639)	All fair	Low: 9% Medium: 86% High: 5%	Low: 9% Medium: 91% High: 0%
RA Benefit: • DAS remission	2 (N=982)	All fair	Low: 9% Medium: 82% High: 9%	Low: 9% Medium: 82% High: 9%
RA Harms: • Serious infection	4 (N=1,215)	All fair	Low: 14% Medium: 82% High: 4%	Low: 18% Medium: 82% High: 0%
RA Harms: • Infusion or injection reaction	4 (N=1,108)	All fair	Low: 9% Medium: 91% High: 0%	Low: 9% Medium: 91% High: 0%

Abbreviations: ACR = American College of Rheumatology; DAS; Disease Activity Scale; HAM-D = Hamilton Rating Scale for Depression; MADRS = Montgomery-Åsberg Depression Rating Scale; MDD = major depressive disorder; N = number; RA = rheumatoid arthritis; RCT = randomized controlled trial.

Most observational studies had been rated fair quality; in two of the exercises, we included a single study that had been rated good quality (Table 7).

Table 7. Summary of observational study risk of bias domain scores, by condition and outcome

Condition Outcome	Number of Studies (Total N of subjects)	Study Quality Ratings	Individual Domain Scores	Reconciled Domain Scores
MDD Harms: • Sexual dysfunction	2 (N=3,154)	Both fair	Low: 14% Medium: 41% High: 45%	Low: 10% Medium: 45% High: 45%
MDD Harms: • Suicidality	2 (N=11,350)	1 fair, 1 good	Low: 27% Medium: 41% High: 32%	Low: 18% Medium: 55% High: 30%
RA Benefits: • ACR20	2 (N=2,461)	Both fair	Low: 4% Medium: 59% High: 36%	Low: 0% Medium: 64% High: 36%
RA Benefit: • ACR70	1 (N=269)	Fair	Low: 5% Medium: 52% High: 43%	Low: 9% Medium: 55% High: 36%
RA Benefits: • DAS remission	1 (N=1,083)	Fair	Low: 18% Medium: 50% High: 32%	Low: 0% Medium: 73% High: 27%
RA Harms: • Serious infection	2 (N=7,695)	1 fair, 1 good	Low: 18% Medium: 50% High: 32%	Low: 9% Medium: 64% High: 27%

Abbreviations: ACR = American College of Rheumatology; DAS = Disease Activity Scale; MDD = major depressive disorder; N = number; RA = rheumatoid arthritis.

A potential limitation of the risk of bias evaluation is that we did not give reviewers the specific criteria that had been used to develop the original individual quality ratings. Thus, they did not know the specific bias risks that would have caused a study to be rated as “fair” quality rather than “good.” Also, reviewers were not told whether limitations inherent in observational study design types had been incorporated into the quality rating that had been recorded for each of the observational studies. However, because two of the observational studies had been rated as “good” quality, our reviewers could have (reasonably) concluded that the original EPC authors were unlikely to have downgraded observational studies based on study design.

Randomized Controlled Trials

In relation to the level of risk of bias of the RCTs included in the exercises, we found substantial agreement among both individuals (AC1=0.67) and pairs of raters (AC1=0.65) (Table 5). The study team had anticipated a high level of agreement among raters because all of the RCT studies had initially been rated fair quality.

With respect to comments from reviewers who found the evaluation of RCT risk of bias to be difficult, five of seven were concerned that the RA studies did not compare individual drugs. However, we had instructed reviewers to treat each of the drugs within a class as equivalent;

thus, some unknown portion of the differences among reviewers' risk of bias scores may be attributable to misinterpretation of these instructions.

In addition, two raters found evaluating risk of bias to be difficult when outcomes were not similarly specified across studies. This concern was expressed about the sexual dysfunction outcome for the MDD second-generation antidepressants study.

Observational Studies

Across the six exercises that included observational studies, agreement among individual reviewers for risk of bias can be considered only slight (AC1=0.11). Agreement improved to fair (AC1=0.22) among pairs of reviewers for the reconciled responses. Even so, this level of agreement did not approach that for RCTs (Table 5).

Comments from those reviewers who thought that scoring risk of bias was difficult chiefly concerned a lack of confidence in the appropriate approach for evaluating this domain for observational studies. One reviewer thought that if just one observational study was being evaluated, then the risk of bias was high. This reviewer's decision is in contrast to the EPC Methods Guide recommendation that the consistency domain score be used to capture the limitation of evidence from just one study.[2] A second reviewer did not know how to evaluate an outcome, in this case sexual dysfunction, when it had been measured differently across studies. Similarly, this concern is discussed in guidance on rating the consistency domain. A third claimed that if included studies had different quality ratings, then the domain score should reflect the lowest contributing element. Finally, a fourth reviewer was not sure how to evaluate observational studies compared with RCTs; this individual did not know whether a high-quality observational study could be considered low risk of bias and chose never to consider such a study as better than medium risk.

Consistency

RCTs

Exercises evaluating MDD and RA outcomes based on RCT studies both included three outcomes with virtually all results in the same direction; two included only results that were not significantly different (Table 8). Two additional MDD outcomes were evaluated through just one study. Two additional RA outcomes included mixed results, including one study that did not report the magnitude of nonsignificant findings. The length of time over which outcomes had been evaluated varied across studies in these two CERs.

Scoring the consistency domain for RCTs across exercises produced moderate agreement among individuals (AC1=0.51); the level of agreement increased to substantial among pairs (AC1=0.70) (Table 5).

Independent reviewers were in poorest agreement in their consistency grade for serious infection from RA medications; 41 percent of the 22 individual reviewers found the studies consistent, 36 percent thought they were inconsistent, and 23 percent rated consistency as unknown (Table 8). Serious infections are a rare adverse event; moreover, study results had not always been in the same direction. Once the reviewer pairs had reconciled their scores, 91 percent of the pairs concluded that the results were consistent.

Reviewers who expressed difficulty in deciding on a grade were not confident about how to rate outcomes that included heterogeneous measures, even if the direction of the results had been the same. That is, they were unsure whether to consider them consistent or not applicable because just one study included each measure.

Table 8. Summary of RCT consistency domain scores, by condition and outcome

Condition Outcome	Number of Studies (Total N of subjects)	Direction of Difference Between Treatments and Significance	Individual Doman Scores	Reconciled Domain Scores
MDD Benefits: • HAM-D response	5 (N=690)	4 of 5 studies: same direction; all studies: NS	Consistent: 91% Inconsistent: 9% Unknown: 0%	Consistent: 100% Inconsistent: 0% Unknown: 0%
MDD Benefits: • HAM-D & MADRS response in elderly	1 (N=108)	1 study	Consistent: 5% Inconsistent: 0% Unknown: 95%	Consistent: 0% Inconsistent: 0% Unknown: 100%
MDD Harms: • Sexual dysfunction	4 (1 open label) (N=904)	All studies: same direction, different measures, some sig	Consistent: 14% Inconsistent: 36% Unknown: 23%	Consistent: 45% Inconsistent: 45% Unknown: 10%
MDD Harms: • Suicidality	1 (N=90)	1 study	Consistent: 5% Inconsistent: 5% Unknown: 91%	Consistent: 0% Inconsistent: 0% Unknown: 100%
MDD Harms: • Nausea	5 (N=689)	4 of 5 studies: same direction; all studies: NS	Consistent: 82% Inconsistent: 14% Unknown: 4%	Consistent: 91% Inconsistent: 9% Unknown: 0%
RA Benefit: • ACR20	5 (N=1,639)	4 of 5 studies: same direction; all studies: NS	Consistent: 10% Inconsistent: 86% Unknown: 4%	Consistent: 9% Inconsistent: 91% Unknown: 0%
RA Benefit: • ACR70	5 (N=1,639)	3 studies in same direction, 1 other direction, 1 data not reported, 4 studies: NS	Consistent: 48% Inconsistent: 48% Unknown: 4%	Consistent: 45% Inconsistent: 55% Unknown: 0%
RA Benefit: • DAS remission	2 (N=982)	Virtually all measures same direction, all studies: NS	Consistent: 82% Inconsistent: 4% Unknown: 14%	Consistent: 91% Inconsistent: 0% Unknown: 9%
RA Harms: • Serious infection	4 (N=1,215)	Mixed direction, all studies: NS	Consistent: 41% Inconsistent: 36% Unknown: 23%	Consistent: 91% Inconsistent: 9% Unknown: 0%
RA Harms: • Infusion or injection reaction	4 (N=1,108)	All same direction. 3 of 4 sig	Consistent: 82% Inconsistent: 14% Unknown: 4%	Consistent: 91% Inconsistent: 9% Unknown: 0%

Abbreviations: ACR = American College of Rheumatology; DAS = Disease Activity Scale; HAM-D = Hamilton Rating Scale for Depression; MADRS = Montgomery-Åsberg Depression Rating Scale; MDD = major depressive disorder; N = number; NS = not significant; RA = rheumatoid arthritis; RCT = randomized controlled trial; sig = statistically significant.

Observational Studies

Exercises evaluating the consistency of outcomes based on observational studies were complicated. Three exercises had results in the same direction (two included indirect evidence—i.e., based on comparisons with a third reference drug); one had mixed results; and two included just one study (Table 9). Agreement was fair among individuals (AC1=0.40) and increased to moderate among pairs (AC1=0.55).

Agreement among reviewers was lowest for the outcomes of sexual dysfunction and suicidality. Both of these outcomes were evaluated through two studies with results in the same direction; however, for each outcome, evidence included one direct comparison and one indirect comparison.

Comments from reviewers indicated that they were concerned that they were unable to judge whether differences in treatments in the indirect comparisons were statistically significant and, therefore, whether the evidence was consistent.

Table 9. Summary of observational study consistency domain scores, by condition and outcome

Condition Outcome	# of Studies (Total N)	Direction of Difference Between Treatments and Significance	Individual Domain Scores	Reconciled Domain Scores
MDD Harms: • Sexual dysfunction	2 (N=3,154)	Both studies: same direction, 1 indirect. 1 sig	Consistent: 50% Inconsistent: 18% Unknown: 32%	Consistent: 64% Inconsistent: 9% Unknown: 27%
MDD Harms: • Suicidality	2 (N=11,350)	Both studies: same direction, 1 indirect. Both: NS	Consistent: 59% Inconsistent: 18% Unknown: 23%	Consistent: 55% Inconsistent: 18% Unknown: 30%
RA Benefits: • ACR20	2 (N=2,461)	Both studies: same direction. Both sig	Consistent: 73% Inconsistent: 14% Unknown: 14%	Consistent: 73% Inconsistent: 0% Unknown: 27%
RA Benefit: • ACR70	1 (N=269	1 study	Consistent: 0% Inconsistent: 5% Unknown: 95%	Consistent: 0% Inconsistent: 0% Unknown: 100%
RA Benefits: • DAS remission	1 (N=1,083)	1 study	Consistent: 0% Inconsistent: 9% Unknown: 91%	Consistent: 0% Inconsistent: 0% Unknown: 100%
RA Harms: • Serious infection	2 (N=7,695)	1 study found difference (HR=1.39), 1 did not (RR=1)	Consistent: 23% Inconsistent: 68% Unknown: 9%	Consistent: 9% Inconsistent: 91% Unknown: 0%

Abbreviations: ACR = American College of Rheumatology; DAS = Disease Activity Scale; HR = hazard ratio; MDD = major depressive disorder; N = number; NS = not significant; RA = rheumatoid arthritis; RR = relative risk; sig = statistically significant

Directness

RCTs

All RCT exercises were limited to direct head-to-head comparisons (Table 10). Generally, outcomes measured ultimate endpoints of interest. Agreement was substantial among individuals (AC1=0.73) and pairs (AC1=0.78) (Table 5). Suicidality is the one exception; data from the original CER had not included a study measuring the endpoint of suicide. Instead, reviewers were asked to consider one study that relied on measures of suicide attempts, ideation, and a scale score. Inter-rater reliability was poorest in relation to this outcome: 55 percent of individual reviewers considered the outcomes direct, whereas 45 percent considered them indirect.

With respect to difficulty of this task, several reviewers expressed concern that they were not always confident of when to consider an outcome direct or indirect. One reviewer expressed confusion but concluded that all evidence with direct links to outcomes should be thought of as direct. A second thought that the meaning of directness was “arguable” but did not provide an explanation.

Table 10. Summary of RCT directness domain scores, by condition and outcome

Condition Outcome	# of Studies (Total N)	Directness Summary	Individual Domain Scores	Reconciled Domain Scores
MDD Benefits: • HAM-D response	5 (N=690)	All head-to-head, outcome measured through 1 scale	Direct: 82% Indirect: 18%	Direct: 82% Indirect: 18%
MDD Benefits: • HAM-D & MADRS response in elderly	1 (N=108)	Head-to-head, outcome measured through 2 scales	Direct: 77% Indirect: 23%	Direct: 82% Indirect: 18%
MDD Harms: • Sexual dysfunction	4 (1 open label) (N=904)	Head-to-head, outcome measured through a variety of measures	Direct: 96% Indirect: 4%	Direct: 91% Indirect: 9%
MDD Harms: • Suicidality	1 (N=90)	Head-to-head, outcome measured through attempts, ideation and scale score	Direct: 55% Indirect: 45%	Direct: 45% Indirect: 55%
MDD Harms: • Nausea	5 (N=689)	All head-to-head	Direct: 100% Indirect: 0%	Direct: 100% Indirect: 0%
RA Benefit: • ACR20	5 (N=1,639)	All head-to head, outcome measured through 1 scale	Direct: 86% Indirect: 14%	Direct: 100% Indirect: 0%
RA Benefit: • ACR70	5 (N=1,639)	All head-to-head, outcome measured through 1 scale	Direct: 86% Indirect: 14%	Direct: 91% Indirect: 9%
RA Benefit: • DAS remission	2 (N=982)	Both head-to-head, outcome measured through 1 scale	Direct: 95% Indirect: 5%	Direct: 100% Indirect: 0%
RA Harms: • Serious infection	4 (N=1,215)	All head-to-head	Direct: 96% Indirect: 4%	Direct: 100% Indirect: 0%
RA Harms: • Infusion or injection reaction	4 (N=1,108)	All head-to-head	Direct: 91% Indirect: 9%	Direct: 91% Indirect: 9%

Abbreviations: ACR = American College of Rheumatology; DAS = Disease Activity Scale; HAM-D = Hamilton Rating Scale for Depression; MADRS = Montgomery-Åsberg Depression Rating Scale; MDD = major depressive disorder; N = number; RA = rheumatoid arthritis; RCT = randomized controlled trial.

Observational Studies

Observational study exercises included direct and indirect comparisons as well as variability in how included studies measured particular outcomes (Table 11). Agreement was moderate among individuals (AC1=0.48) and substantial among pairs (AC1=0.78) (Table 5). The higher level of agreement among pairs was related primarily to agreement increasing to 100 percent among pairs for the four RA outcomes, which were more clearly ultimate endpoints of interest. Agreement among pairs was less than 100 percent for MDD outcomes: sexual dysfunction (one direct comparison including various outcome measures and one indirect comparison) and suicidality (measured through nonfatal suicidal behavior, nonfatal self harm, and completed suicides).

In relation to these two outcomes, reviewers who expressed difficulty were unsure of what to do when the evidence is mixed. That is, the evidence base might include studies with direct and indirect comparisons, or it might include studies with ultimate and intermediate measures of outcomes or heterogeneous outcome measures (and, of course, it might have all of these complications).

Table 11. Summary of observational study directness domain scores, by condition and outcome

Condition Outcome	# of Studies (Total N)	Directness Summary	Individual Domain Scores	Reconciled Domain Scores
MDD Harms: - Sexual dysfunction	2 (N=3,154)	1 head-to-head comparison, 1 indirect comparison, outcome measured through scales and specific symptoms	Direct: 64% Indirect: 36%	Direct: 64% Indirect: 36%
MDD Harms: - Suicidality	2 (N=11,350)	1 head-to-head comparison, 1 indirect comparison, outcome measured through suicide and suicidal behaviors	Direct: 59% Indirect: 41%	Direct: 64% Indirect: 36%
RA Benefits: - ACR20	2 (N=2,461)	Both head-to-head comparisons, outcome measured through 1 scale	Direct: 82% Indirect: 18%	Direct: 100% Indirect: 0%
RA Benefit: - ACR70	1 (N=269)	1 head-to-head, outcome measured through 1 scale	Direct: 81% Indirect: 19%	Direct: 100% Indirect: 0%
RA Benefits: - DAS remission	1 (N=1,083)	1 head-to-head, outcome measured through 1 scale	Direct: 86% Indirect: 14%	Direct: 100% Indirect: 0%
RA Harms: - Serious infection	2 (N=7,695)	Both head-to-head	Direct: 91% Indirect: 9%	Direct: 100% Indirect: 0%

Abbreviations: ACR = American College of Rheumatology; DAS = Disease Activity Scale; MDD = major depressive disorder; N = number; RA = rheumatoid arthritis.

Precision

RCTs

In relation to outcomes evaluated through RCTs, only one of the 10 exercises provided reviewers with an estimate of precision based on the results from a meta-analysis (Table 12). Two exercises were limited to one study. In both, although differences between drugs for MDD were said to be statistically significant, the authors of the original studies had not presented confidence intervals, and only one article had included outcome rates for both arms of the study. Three exercises included studies of differences that were all not statistically significant, and four included results that were mixed. Inter-rater reliability agreement based on RCT evidence was lowest for this domain: fair among individual raters (AC1=0.23) but moderate (AC1=0.47) after results were reconciled among pairs (Table 5).

Reviewers expressed difficulty in evaluating this domain for the exercises that did not include meta-analyses; they thought that more guidance was needed for that circumstance. Some were concerned that they did not know the level of precision that was required to evaluate an estimate as being precise. In particular, even if an estimate had reported a statistically significant difference between drugs or drug classes, some reviewers expressed difficulty in making determinations without having confidence intervals to evaluate. One reviewer thought that the guidance alluded to the notion that nonsignificant differences could be precise under some circumstances but did not know how to make that determination. Also, reviewers raised questions about how to approach scoring precision when an outcome was evaluated through different measures; they were unsure of whether they should consider the measures in combination or evaluate them separately. Lastly, some reviewers found the scoring difficult when the outcome was assessed through one study and wondered whether the domain should be considered inapplicable in those cases.

Table 12. Summary of RCT precision domain scores, by condition and outcome

Condition Outcome	# of Studies (Total N)	Difference Between Treatments	Individual Domain Scores	Reconciled Domain Scores
MDD Benefits: • HAM-D response	5 (N=690)	Meta-analysis result: 1.03; 95% CI, 0.92 to 1.16	Precise: 82% Imprecise: 18%	Precise: 82% Imprecise: 18%
MDD Benefits: • HAM-D & MADRS response in elderly	1 (N=108)	2 measures, p-value for both: <0.05, rates and CI not provided	Precise: 50% Imprecise: 50%	Precise: 55% Imprecise: 45%
MDD Harms: • Sexual dysfunction	4 (1 open label) (N=904)	Various measures, 1: p<0.05, all others: p=NS; rates provided in 3 of 4 but no CI	Precise: 18% Imprecise: 82%	Precise: 0% Imprecise: 100%
MDD Harms: • Suicidality	1 (N=90)	p=0.026, rate provided, but no CIs	Precise: 18% Imprecise: 82%	Precise: 18% Imprecise: 82%
MDD Harms: • Nausea	5 (N=689)	All studies NS; 4 of 5: rate provided but no CIs	Precise: 45% Imprecise: 55%	Precise: 45% Imprecise: 55%
RA Benefit: • ACR20	5 (N=1,639)	Mix of endpoints: 3 of 5: p<0.05; at 12 months: 2 of 3: p=NS, rates provided, but no CIs	Precise: 18% Imprecise: 82%	Precise: 0% Imprecise: 100%
RA Benefit: • ACR70	5 (N=1,639)	4 of 5: p=NS; 3 of 4: rates provided, but no CIs	Precise: 19% Imprecise: 81%	Precise: 0% Imprecise: 100%
RA Benefit: • DAS remission	2 (N=982)	All outcomes: p=NS; rates for each treatment close, but no CIs	Precise: 45% Imprecise: 55%	Precise: 64% Imprecise: 36%
RA Harms: • Serious infection	4 (N=1,215)	All: NS, no CIs	Precise: 18% Imprecise: 82%	Precise: 36% Imprecise: 64%
RA Harms: • Infusion or injection reaction	4 (N=1,108)	3 of 4: p<0.05, rates provided, but no CIs	Precise: 77% Imprecise: 23%	Precise: 100% Imprecise: 0%

Abbreviations: ACR = American College of Rheumatology; CI = confidence interval; DAS = Disease Activity Scale; HAM-D = Hamilton Rating Scale for Depression; MADRS = Montgomery-Åsberg Depression Rating Scale; MDD = major depressive disorder; N = number; NR = not reported; NS = not significant; RA = rheumatoid arthritis; RCT = randomized controlled trial.

Observational Studies

Agreement among raters on their evaluation of the precision of observational studies was fair for individuals (AC1=0.31) and for pairs (AC1=0.38) (Table 5). Several examples of exercises with poor agreement are instructive. In the exercise evaluating sexual dysfunction, 36 percent of individual reviewers thought the evidence was precise, whereas 64 percent thought it was imprecise (Table 13). This outcome was evaluated through two different measures—a direct comparison with significant results but no confidence intervals presented and an indirect comparison with overlapping confidence intervals for each of the two drugs being compared and a third drug (Table 13). In the exercise evaluating ACR20, 59 percent of independent reviewers through the evidence was precise whereas 41 percent thought it was imprecise. This exercise included statistically significant findings from two studies; the first presented p-values but no point estimates for differences between the drug classes, and the second presented confidence intervals in an adjusted analysis.

Comments from raters concerning the difficulty of evaluating the precision of observational studies were similar to those for RCTs.

Table 13. Summary of observational study precision domain scores, by condition and outcome

Condition Outcome	# of Studies (Total N)	Difference Between Treatments	Individual Domain Scores	Reconciled Domain Scores
MDD Harms: - Sexual dysfunction	2 (N=3,154)	1 direct comparison, multiple measures, all sig and in same direction, no CIs; 1 indirect comparison, NS, overlapping CIs	Precise: 36% Imprecise: 64%	Precise: 45% Imprecise: 55%
MDD Harms: - Suicidality	2 (N=11,350)	1 indirect comparison: NS, overlapping CIs 1 direct nested case-control adjusted comparison: 2 measures, both NS, CIs reported	Precise: 32% Imprecise: 68%	Precise: 27% Imprecise: 73%
RA Benefits: - ACR20	2 (N=2,461)	1 study: sig difference, no data, p-values only, 1 adjusted study: sig diff, CI reported	Precise: 59% Imprecise: 41%	Precise: 45% Imprecise: 55%
RA Benefit: - ACR70	1 (N=269)	Difference reported through p-value only, NS, no results data	Precise: 5% Imprecise: 95%	Precise: 0% Imprecise: 100%
RA Benefits: - DAS remission	1 (N=1,083)	Adjusted analysis: sig, CIs presented, matched pairs analysis: sig, CIs not reported	Precise: 55% Imprecise: 45%	Precise: 82% Imprecise: 18%
RA Harms: - Serious infection	2 (N=7,695)	Adjusted analyses in each study, 1 sig, 1 NS, CIs presented for both	Precise: 14% Imprecise: 86%	Precise: 36% Imprecise: 64%

Abbreviations: ACR = American College of Rheumatology; CI = confidence interval; DAS = Disease Activity Scale; MDD = major depressive disorder; N = number; NS = not significant; RA = rheumatoid arthritis; sig = statistically significant.

Strength of Evidence

Inter-rater reliability agreement was slight among individual reviewers (AC1=0.20) but improved to fair (AC1=0.24) after results were reconciled among pairs (Table 5). On average, almost 20 percent of reviewers reported that the exercise of grading the overall strength of evidence was difficult or very difficult.

We considered whether reviewers' agreement on the strength of evidence grade was greater when they completed exercises that included only RCTs (four exercises) than when they completed exercises that included both RCTs and observational studies (all 10 exercises). Inter-rater reliability did not improve appreciably; agreement was fair among individual reviewers (AC1=0.22) and reconciled pairs (AC1=0.30) for the four exercises that include only RCTs.

Table 14 presents the grading results for each exercise. Even after grades were reconciled with partners, many results varied widely. At the extreme, reconciled grades for the MDD harm of nausea included all four possibilities (high, moderate, low, and insufficient), as did the RA benefit of ACR70 scores and the RA harm of severe infection.

Table 14. Strength of evidence grades, by condition and outcome

Condition Outcome	Individual SOE Grades	Reconciled SOE Grades
MDD Benefits: • HAM-D response	High: 23% Moderate: 64% Low: 14% Insufficient: 0%	High: 18% Moderate: 73% Low: 9% Insufficient: 0%
MDD Benefits: • HAM-D & MADRS response in elderly	High: 0% Moderate: 5% Low: 45% Insufficient: 50%	High: 0% Moderate: 0% Low: 45% Insufficient: 55%
MDD Harms: • Sexual dysfunction	High: 0% Moderate: 18% Low: 64% Insufficient: 18%	High: 0% Moderate: 9% Low: 82% Insufficient: 9%
MDD Harms: • Suicidality	High: 5% Moderate: 18% Low: 36% Insufficient: 41%	High: 0% Moderate: 9% Low: 45% Insufficient: 45%
MDD Harms: • Nausea	High: 19% Moderate: 62% Low: 10% Insufficient: 10%	High: 18% Moderate: 55% Low: 9% Insufficient: 18%
RA Benefit: • ACR20	High: 0% Moderate: 18% Low: 59% Insufficient: 23%	High: 0% Moderate: 18% Low: 55% Insufficient: 27%
RA Benefit: • ACR70	High: 5% Moderate: 19% Low: 62% Insufficient: 14%	High: 9% Moderate: 18% Low: 55% Insufficient: 18%
RA Benefit: • DAS remission	High: 0% Moderate: 19% Low: 52% Insufficient: 29%	High: 0% Moderate: 9% Low: 55% Insufficient: 36%
RA Harms: • Serious infection	High: 9% Moderate: 18% Low: 45% Insufficient: 27%	High: 9% Moderate: 27% Low: 45% Insufficient: 18%
RA Harms: • Infusion or injection reaction	High: 27% Moderate: 55% Low: 18% Insufficient: 0%	High: 18% Moderate: 73% Low: 9% Insufficient: 0%

Abbreviations: ACR = American College of Rheumatology; DAS = Disease Activity Scale; HAM-D = Hamilton Rating Scale for Depression; MADRS = Montgomery-Åsberg Depression Rating Scale; MDD = major depressive disorder; RA = rheumatoid arthritis; SOE = strength of evidence.

We explored patterns among reviewer pairs in the relationship between domain scores and strength of evidence grades for two exercises with very divergent conclusions. Table 15 presents reconciled domain scores and strength of evidence grades for each reviewer pair from the exercises on MDD harm of nausea (five studies, all RCTs) and the RA benefit of ACR70 (six studies, five of which were RCTs).

Table 15. Detail of reconciled pair assessments for two exercises with poor agreement across pairs: antidepressant harms (nausea)[a] and rheumatoid arthritis benefit (ACR70)

Reviewer Pair[b]	Risk of Bias[c]		Consistency[d]		Directness[e]		Precision[f]		Strength of Evidence[g]
	RCTs	Obs	RCTs	Obs	RCTs	Obs	RCTs	Obs	
Nausea 1.	Low		Consistent		Direct		Imprecise		High
Nausea 2.	Low		Consistent		Direct		Precise		High
Nausea 3.	Medium		Consistent		Direct		Precise		Moderate
Nausea 4.	Medium		Consistent		Direct		Imprecise		Moderate
Nausea 5.	Medium		Consistent		Direct		Precise		Moderate
Nausea 6.	Medium		Consistent		Direct		Imprecise		Moderate
Nausea 7.	Medium		Consistent		Direct		Imprecise		Moderate
Nausea 8.	Medium		Consistent		Direct		Precise		Moderate
Nausea 9.	Medium		Consistent		Direct		Imprecise		Low
Nausea 10.	Medium		Consistent		Direct		Imprecise		Insufficient
Nausea 11.	Medium		Inconsistent		Direct		Imprecise		Insufficient
ACR70 1.	Low	High	Inconsistent	NA	Direct	Direct	Imprecise	Imprecise	High
ACR70 2.	Medium	Medium	Consistent	NA	Direct	Direct	Imprecise	Imprecise	Moderate
ACR70 3.	Medium	Medium	Inconsistent	NA	Indirect	Direct	Imprecise	Imprecise	Moderate
ACR70 4.	Medium	High	Inconsistent	NA	Direct	Direct	Imprecise	Imprecise	Low
ACR70 5.	Medium	Medium	Inconsistent	NA	Direct	Direct	Imprecise	Imprecise	Low
ACR70 6.	Medium	High	Inconsistent	NA	Direct	Direct	Imprecise	Imprecise	Low
ACR70 7.	Medium	Medium	Inconsistent	NA	Direct	Direct	Imprecise	Imprecise	Low
ACR70 8.	Medium	High	Consistent	NA	Direct	Direct	Imprecise	Imprecise	Low
ACR70 9.	Medium	Medium	Consistent	NA	Direct	Direct	Imprecise	Imprecise	Low
ACR70 10	Medium	Low	Consistent	NA	Direct	Direct	Imprecise	Imprecise	Insufficient
ACR70 11.	Medium	Medium	Consistent	NA	Direct	Direct	Imprecise	Imprecise	Insufficient

[a] This exercise included only RCTs.

[b] The number assigned to each review pair is arbitrary and does not represent the same pair for the nausea and ACR70 evaluations.

[c] Risk of bias scoring choices were low, medium or high

[d] Consistency scoring choices were consistent, inconsistent, or unknown (NA)

[e] Directness scoring choices were direct or indirect

[f] Precision scoring choices were precise and imprecise

[g] Strength of evidence grading choices were high, medium, low, or insufficient

Abbreviations: ACR =, American College of Rheumatology; NA = consistency score not applicable (single study); Obs = observational studies; RCTs = randomized controlled trials.

The antidepressant harm outcome of nausea evaluation consisted of all fair-quality RCTs; each had a relatively small sample size. In four of the studies, a larger percentage of patients experienced nausea in the paroxetine arm but none of the differences were statistically significant (presented only through p-values). Only the two pairs that scored RCT risk of bias as low concluded that the strength of evidence was high. Five of the 11 pairs of reviewers had all of the same domain score assessments but they differed in their strength of evidence grade; moderate (three pairs), low (one pair), or insufficient (one pair). This was likely related to different views of the appropriate strength of the evidence grade when the evidence is imprecise; other scores among these pairs were medium risk of bias, consistent and direct. Three other of the 11 pairs, evaluated the evidence as medium risk of bias, consistent, direct but precise, and all of these pairs graded the strength of evidence as moderate.

The ACR70 evaluation is a good example of the complexity of the decision faced by reviewers in an evaluation that included both RCTs and observational studies. In this exercise, four of five trials and the one included large observational study did not find significant differences between treatments. One of the RCTs and the observational study did not report the percentage of patients in each arm who achieved the outcome. All studies were considered fair quality and outcomes were evaluated after different lengths of time. Statistical significance was presented only through p-values.

As was found in the nausea exercise, a conclusion of high strength of evidence was made only when RCT risk of bias was assessed as low. Patterns of differences in ACR70 strength of evidence grading were not explained by domain scores. For example, 3 of the 11 pairs had all of the same domain score assessments but reached different strength of evidence grades (moderate, low, and insufficient).

Comments from reviewers indicated that higher grades were associated with confidence in a conclusion that the evidence in this exercise summed to no difference in effect. Although distinguishing the reasons behind grades of low versus insufficient is difficult, both of these groups of reviewers (the majority of participants) seemed more reluctant to give a higher grade to findings that appeared to show no difference.

Predictions of Strength of Evidence Grades From Domain Scores

We used two multinomial logistic regression analyses to examine the relationship between independent reviewers' strength of evidence grades and domain scores (Table 16). Based on results from the six exercises that included RCTs and observational studies (Model 1), reviewers who considered the evidence from observational studies as being low risk of bias (vs. high) were more than 500 percent more likely (i.e., OR=6.77) to grade strength of evidence as insufficient (vs. moderate or high). Reviewers who considered the evidence from observational studies as being consistent (vs. unknown consistency) were 86 percent less likely (i.e., OR=0.14) to grade strength of evidence as insufficient (vs. moderate or high). By contrast, reviewers who considered the evidence from RCTs as being precise (vs. imprecise) were 75 percent less likely (i.e., OR=0.25) to grade strength of evidence as low (vs. moderate or high).

Based on results from all 10 exercises (Model 2 in Table 16), the reviewers were more likely to grade strength of evidence as insufficient or low (vs. moderate or high) if the exercise included observational studies, controlling for RCT domain scores. Also, reviewers' strength of evidence grades were significantly less likely to be insufficient if results from RCT evidence were considered either consistent or inconsistent (vs. unknown consistency) or precise (vs. imprecise). They were less likely to grade strength of evidence as low (vs. moderate or high) if the evidence from RCTs were considered consistent (vs. unknown consistency) or precise (vs. imprecise).

Table 16. Prediction of strength of evidence grade by domain scores: odds ratio from multinomial logistic regression results

Domain/ Strength of Evidence	Study Design	Domain Score	Strength of Evidence Model 1		Strength of Evidence Model 2	
			vs. Moderate or High		vs. Moderate or High	
			Insufficient	Low	Insufficient	Low
Risk of bias (vs. high)	RCT	Medium	9.70	3.41	2.52	2.78
		Low	1.00	0.55	0.34	0.60
	Observational	Medium	1.76	1.13	NA	NA
		Low	6.77[a]	1.69	NA	NA
Consistency (vs. unknown or NA)	RCT	Consistent	0.18	1.33	0.04[a]	0.12[a]
		Inconsistent	0.90	3.14	0.18[a]	0.36
	Observational	Consistent	0.14[a]	0.32	NA	NA
		Inconsistent	1.99	0.77	NA	NA
Directness (vs. indirect)	RCT	Direct	0.89	0.45	0.93	0.65
	Observational	Direct	0.79	0.55	NA	NA
Precision (vs. imprecise)	RCT	Precise	0.19	0.25[a]	0.25[a]	0.39[a]
	Observational	Precise	0.26	0.56	NA	NA
	With observational studies in exercise (vs. only RCT studies included in exercise)		NA	NA	4.10[a]	6.02[a]

[a]p<0.05

Abbreviations: NA = not applicable; RCT = randomized controlled trials.

Predictions of Level of Agreement in Strength of Evidence Grades From Level of Agreement in Domain Scores

We estimated two logistic regression equations (Models 3 [with observational studies] and 4 [only RCTs]) to examine the association between independent reviewers' agreement on domain scores and their eventual agreement on strength of evidence grade (Table 17). Across all analyses, we did not find that agreement on any of the domain scores predicted agreement on the strength of evidence grade. Taking precision scores for observational studies as one example (in Model 3), reviewers who agreed (relative to not agreeing on the score) were 87 percent more likely (OR=1.87) of agreeing on the strength of evidence grade, but this result was not statistically significant. Looking at only agreement on RCT evidence (in Model 4), reviewers who agreed on RCT risk of bias (relative to not agreeing on the score) were 51 percent more likely (OR=1.51) of agreeing on the strength of evidence grade, but again, this result was not statistically significant.

Table 17. Prediction of agreement on strength of evidence grade by agreement on domain scores: Odds ratios from logistic regression results

Domain	Study Design	Agree Strength of Evidence (vs. not) Model 3	Agree Strength of Evidence (vs. not) Model 4
Risk of bias	RCT	1.56	1.51
	Observational	0.68	NA
Consistency	RCT	1.26	1.71
	Observational	0.91	NA
Directness	RCT	0.84	0.76
	Observational	1.02	NA
Precision	RCT	1.36	1.44
	Observational	1.87	NA
	No observational studies in exercise	NA	1.00

No results were statistically significant (p<0.05)

Abbreviations: NA = not applicable; RCT = randomized controlled trials.

Subgroup Analyses of Independent Reviewer Assessments

We reviewed independent reviewer agreement in domain scores and strength of evidence grades by condition (i.e., MDD or RA), type of outcome (i.e., benefit or harm), reviewer academic training (i.e., physician or nonphysician), and years of experience in conducting systematic or comparative effectiveness reviews and evaluating strength of evidence. Across all subgroups, agreement on the strength of evidence grade was lower than for virtually all domain-specific scores.

With respect to condition (Table 18), reviewers' agreement on seven of eight domain scores was greater (shown in bold) for RA than MDD. Two differences between MDD and RA (consistency and directness for observational studies) were statistically significant. However, agreement in RA domain scoring did not correspond to greater agreement for the RA strength of evidence grade; the level of agreement was nearly identical.

Table 18. Percentage agreement among independent reviewers, by condition (MDD vs. RA) and outcome (benefits vs. harms)

Domain/Strength of Evidence	Study Design	MDD	RA	Benefits	Harms
Risk of Bias	RCTs	82%	**86%**	82%	**86%**
	Observational	43%	**53%**	**54%**	45%
Consistency	RCTs	**80%**	73%	**80%**	73%
	Observational	55%*	**82%***	**86%***	59%*
Directness	RCTs	82%	**92%**	85%	**88%**
	Observational	62%*	**85%***	**83%**	71%
Precision	RCTs	70%	**72%**	70%	**72%**
	Observational	66%	**74%**	70%	**73%**
Strength of Evidence		**56%**	54%	**58%**	52%

* p<0.05, based on individual-level data and controlling for repeated measures. (N = 22)

Note: bolded score denotes greater agreement across reviewers.

Abbreviations: MDD = major depressive disorder; RA = rheumatoid arthritis; RCT = randomized controlled trials.

By type of outcome (Table 18), half of the eight comparisons showed greater agreement for benefits; only agreement about consistency for observational studies was significantly different between benefits and harms. Reviewers agreed 58 percent of the time on strength of evidence grades for exercises examining benefits and 52 percent of the time for exercises examining harms.

Of 22 reviewers, eight were physicians by academic training; the remainder represented a range of social science and methodologic backgrounds (Table 19). Of the eight comparisons of agreement on domain scores, six showed greater agreement among those reviewers with professional backgrounds other than medicine. However, the one significant result—agreement on scoring risk of bias for RCTs—showed greater agreement among physician than nonphysician reviewers. Like the analysis by condition, greater agreement across domain scoring among nonphysicians did not correspond with greater agreement in strength of evidence grading (58 percent of physicians and 53 percent of nonphysicians).

Table 19. Percentage agreement among independent reviewers, by reviewer academic training, experience in conducting systematic or comparative effectiveness reviews, and experience evaluating strength of evidence

Domain/Strength of Evidence	Study Design	Physician Academic Training	Non-Physician Academic Training	>= 10 Reviews	<10 Reviews	>=5 Evaluations of SOE	<5 Evaluations of SOE
Risk of Bias	RCTs	**96%***	77%*	**88%**	79%	80%	**86%**
	Observational	**79%**	76%	**56%**	41%	45%	**52%**
Consistency	RCTs	84%	**88%**	**78%**	74%	70%	**80%**
	Observational	68%	**72%**	70%	**76%**	62%	**78%**
Directness	RCTs	42%	**54%**	**88%**	84%	**90%**	85%
	Observational	71%	**73%**	**82%**	70%	**79%**	76%
Precision	RCTs	69%	**82%**	69%	**72%**	**73%**	69%
	Observational	65%	**75%**	**77%***	63%*	**79%**	67%
Strength of Evidence		**58%**	53%	50%	**61%**	46%	**59%**

*p<0.05, based on individual-level data and controlling for repeated measures. (Reviewer academic training: N = 22; Number of reviews and number of times evaluating strength of evidence: N = 21)

Note: bolded score denotes greater agreement across reviewers.

Abbreviations: RCT = randomized controlled trials; SOE = strength of evidence.

Reviewer experience conducting reviews and evaluating strength of evidence was available for 21 reviewers. Experience conducting reviews ranged from two reviews (two reviewers) to 25 or more reviews (five reviewers); 10 reviewers had conducted 10 or more reviews. More experienced reviewers showed greater agreement for six of eight domain scores (only scoring of precision of observational studies was statistically significant). Greater agreement across domain scoring among more experienced reviewers corresponded with poorer agreement in strength of evidence grading (50 percent of more experienced reviewers and 61 percent of less experienced reviewers.)

Reviewer experience evaluating strength of evidence ranged from none (two reviewers) to 10 or more times (four reviewers); seven reviewers had evaluated strength of evidence five or more times. Reviews with greater experience in evaluating strength of evidence showed greater agreement across half of the eight domain scores (none of the differences were statistically significant). Like the analysis of experience in conducting reviews, greater experience in evaluating strength of evidence corresponded with poorer agreement in strength of evidence grading (46 percent of more experienced graders and 59 percent of less experienced graders).

Reconciliation Across Reviewer Pairs

We received responses from 8 of the 11 pairs describing the approach each of the reviewers had used independently to assign overall strength of evidence grades. They generally described their approach as qualitative, reflecting the lack of meta-analytic results data in 9 of the 10 exercises that would have lent themselves to quantitative analyses. The methods approach for qualitative assessment differed across reviewers. Two pairs included independent reviewers who used a different approach: in one pair, one reviewer used the EPC's own approach of assessing the direction and size of effect separately, while the other followed guidance in Owens et al.[2]; in the second, one reviewer used the GRADE approach and the second used the guidance in Owens et al.[2] Other pairs of reviewers used the same approach as each other.

The approaches were described as in various ways: a subjective judgment based on a mix of quantitative and qualitative analyses; a qualitative approach based on guidance in Owens et al.;[2] the EPC's own approach; the EPCs own qualitative approach guided by GRADE, Owens et al.,[2] and US Preventive Services Task Force guidance; an approach based on Owens et al.[2] with rules for assigning particular grades, such as high strength of evidence limited to outcomes with "perfect" domain scores (i.e., low risk of bias, consistent, direct, and precise); and lastly, the EPC's own approach first and then consideration of guidance in Owens et al.[2] We did not conduct subgroup analyses based on the approach used to assign grades because we lacked sufficient data to sort approaches into meaningful categories.

Only one pair of reviewers used a third party to adjudicate differences in their domain scores or strength of evidence grades. We did not conduct subgroup analyses based on this distinction because with only one EPC using adjudication, we would have been unable to conclude that any differences in results were based on the pair's adjudication per se or other unmeasured characteristics of the reviewer pair.

The majority of individually determined domain scores were the same for both partners; i.e., these scores did not change during the reconciliation process (Table 20). Taking risk of bias as the example, summing across all 10 outcomes, pairs agreed on 82 percent of the RCT risk of bias scores (top row of Table 20); of the remaining scores, they reconciled 9 percent of their scores to be better (lower risk of bias) and 9 percent to be worse (higher risk of bias) than either or both of their individual scores. By contrast, for observational studies precision, they agreed on 67 percent of domain scores and reconciled 18 percent and 15 percent, respectively, to be better (precise) or worse (imprecise) in contrast to either or both of their individual scores.

Table 20. Mean percentage change in domain scores and the strength of evidence grade following reconciliation

Domain/Strength of Evidence	Study Design	Agree (No change)	Better (Higher score)	Worse (Lower score)
Risk of bias	RCT	82%	9%	9%
	Observational	50%	24%	26%
Consistency	RCT	69%	21%	10%
	Observational	77%	11%	12%
Directness	RCT	82%	11%	7%
	Observational	65%	27%	8%
Precision	RCT	67%	15%	18%
	Observational	67%	18%	15%
Strength of Evidence	Overall combined	46%	25%	30%

Results for domain scores for RCT studies and strength of evidence are based on 11 pairs evaluating 10 outcomes (N = 110).
Results for domain scores on observational studies are based on 11 pairs evaluating 6 outcomes (N = 66).
Abbreviations: RCT = randomized controlled trial.

Agreement for domain scores ranged from a low of 50 percent to a high of 82 percent. For changes in scores that needed to be reconciled, we found no consistent pattern toward either better or worse scores. One exception may be of interest: the reconciled scores for directness for RCTs and observational studies were both in the direction of "better" scores (going from indirect to direct).

Initial agreement across pairs of reviewers—46 percent—was poorer for overall strength of evidence grades than it was for any of the domain scores. Of the remaining overall grades, about half were reconciled to be better (25 percent) and the remainder were reconciled worse (30 percent).

The pattern of reconciliation to better or worse overall strength of evidence grades was not generally related to the grades that were being reconciled. For example, among the 22 assessments where one reviewer in the pair originally assessed the grade as low and the other reviewer assessed the grade as insufficient, 11 were reconciled to low and 11 were reconciled to insufficient (data not shown). Similarly, among the 11 assessments where one reviewer in the pair originally assessed the grade as high and the other reviewer assessed the grade as moderate, five were reconciled to high, five to moderate, and one to low (data not shown). In contrast, among the seven assessments where one reviewer in the pair originally assessed the grade as moderate and the other reviewer assessed the grade as insufficient, five were reconciled to insufficient (data not shown).

Discussion

Overview of the Project

We tested the inter-rater reliability of the AHRQ approach to evaluating strength of evidence. We asked two independent reviewers to provide individual domain scores and strength of evidence grades for 10 exercises and to reconcile their decisions with a partner, reflecting the process that EPCs typically use. Twenty-two reviewers from nine EPCs and AHRQ participated. We solicited comments from reviewers throughout the process to obtain insights concerning their approach to completing the exercise, the evaluations that they considered difficult, and the resulting choices that they made.

Our approach expanded in several ways on a similar empirical study conducted by members of the GRADE working group in the early 2000s.[3] Like the GRADE study, we assessed agreement in the overall grade. However, we began one step earlier in the process and evaluated agreement on the criteria (i.e., domains) that the AHRQ approach requires to decide on the overall grade; these domains are risk of bias, consistency, directness, and precision. Whereas GRADE used data from one review on one disease condition, we used data from two completed comparative effectiveness reviews (CERs) to enhance the generalizability of our findings; these CERs covered different health conditions, a range of benefits, and minor to severe harms.

We sought to determine the level of consistency (i.e., agreement) in domain scores (Key Question 1) and strength of evidence grades (Key Question 2) across reviewers, both when reviewers were working independently and after they had reconciled their decisions with a partner. The 10 exercises included a mix of RCTs and observational studies. They also posed different evaluation challenges, such as heterogeneous outcome measures, indirect comparisons, precision of estimates presented through p-values only, and inconsistent results across studies. Only one exercise lent itself to a quantitative evaluation of pooled estimates through the inclusion of a meta-analysis.

Key Question 1. Principal Findings: Domain Scores

Independent Reviewers

The level of independent reviewer inter-rater agreement for domain scores varied considerably from substantial for RCT risk of bias (AC1 = 0.67) and directness (AC1 = 0.73) to slight for observational study risk of bias (AC1 = 0.11). Agreement on all other domains was either moderate or fair. The high level of consistency of agreement for both these RCT assessments was most likely related to the straightforward nature of the data in the evaluations across exercises, resulting in similar reviewer judgments. All the RCT risk of bias assessments involved studies identified as fair quality; the RCT directness assessments were all head-to-head comparisons, with six of the 10 exercises including outcomes measuring ultimate endpoints of interest through one scale.

In contrast, the risk of bias assessment for observational studies was more problematic. This relates most likely to reviewers' not receiving sufficient guidance concerning the criteria that the project CER teams had originally used for determining the "quality" of the studies, requiring reviewers to use their own judgment criteria to interpret the meaning of the scores. Some, but not all reviewers, considered the evaluation of observational studies to begin as high risk of bias, and even in the two exercises that included one good-quality and one fair-quality study, one-third of reviewers evaluated the risk of bias as high. Comments from reviewers support the conclusion that they lacked confidence in the appropriate approach for making the risk of bias assessment for observational studies.

Inter-rater agreement was greater for RCTs than observational studies, with the exception of the precision domain. For both RCTs and observational studies, agreement on precision was only fair and reviewers raised similar concerns with respect to both types of study designs. They expressed a desire for greater guidance to direct their judgments when they could not rely on a quantitative synthesis through a meta-analysis and were faced with such problems as statistical significance expressed through p-values but not confidence intervals, a variety of differently measured outcomes, and nonsignificant findings.

Reflecting these concerns related to scoring the precision domain, of all of the domain scoring decisions, reviewers were most likely to have considered scoring precision to be difficult or very difficult (18 percent of RCT and 15 percent of observational study decisions). Few reviewers considered scoring other domains to be difficult. For this reason, reviewers' judgments of difficulty were not especially informative with respect to levels of agreement on domain scores.

Reviewer Pairs

Agreement on domain scores for reconciled reviewer pairs was as good as or better than it was for individual independent reviewers. Agreement on three of the four RCT scores was substantial (risk of bias, consistency, and directness). Agreement on precision domain scores was poorer than for the other domains, but it improved from fair to moderate for pairs. Agreement on domain scores for observational studies across reconciled pairs also improved in all domains except precision, but agreement was substantial only for directness.

The direction of change in scores when a pair of reviewers had to settle a difference (i.e., had to reconcile their original scores) was inconsistent across domains. That is, they were reconciled to be "better" or "worse" in no obvious pattern. We had insufficient data to evaluate the effect of using consensus discussion versus a third party adjudicator on level of agreement across pairs.

Based on these findings, we conclude that the reconciliation process is a critical step in domain scoring. We do not have sufficient data to report on differences in agreement based on methodological approaches that the individuals in each of the pairs might have followed. Nevertheless, in relation to RCT data at least, all approaches that these reviewers used, when combined with a reconciliation process, yielded a high level of similar assessment decisions. Achieving a high level of consistency in assessing domains for observational studies is more problematic than for RCTs. However, even for these types of studies, the reconciliation process had a positive effect.

Key Question 2. Principal Findings: Strength of Evidence Grades

Independent Reviewers

Agreement on independent reviewer strength of evidence grades overall was generally poorer than for domain scores. Across all 10 exercises, inter-rater reliability agreement for overall strength of evidence was slight. The level of agreement based on results from a subsample of four exercises that evaluated only evidence from RCTs was fair. Similarly, overall agreement in the GRADE study of RCT-only evidence had been fair.

When evidence was limited to RCT studies, better strength of evidence grades of moderate or high (compared with both insufficient and low) were related to RCT domain scores' being considered consistent and precise. Because all RCT studies were presented as fair quality and were head-to-head trials, it is not surprising that the risk of bias and consistency domains were not predictors of the final strength of evidence grade.

The inclusion of observational studies, in addition to RCTs, in an exercise was a strong predictor of a poorer strength of evidence grade —namely, either insufficient or low versus moderate or high. Comments from reviewers seemed to indicate that they found the body of evidence to be more problematic to assess when these additional study designs were included. They expressed uncertainty concerning how to integrate findings from observational studies to support findings from RCTs.

Looking more closely at the relationship between specific observational studies domain scores and a strength of evidence grade of insufficient or low, we found that observational study evidence being considered low risk of bias was significantly related to strength of evidence being graded as insufficient (but not low) versus moderate or high. This counterintuitive finding may have been driven by the data in the two exercises with good-quality observational studies; the findings from these observational studies conflicted with the findings from the available RCT evidence. By considering the observational studies as well as the RCT data, reviewers might reasonably have concluded that they could not reach a conclusion about the body of evidence. Likewise, consistency of observational and RCT study findings was positively related to the strength of evidence grade being moderate or high versus insufficient. This finding would seem to reflect reviewers' consideration of observational studies as secondary evidence that supports RCT evidence when the direction of the findings are clear and not in conflict with what was found through RCTs. In contrast, we were unable to find analogous patterns that would help explain strength of evidence grades of low versus moderate or high.

In relation to agreement on strength of evidence grades, we found that neither agreement on domain scores nor agreement about the level of difficulty that reviewers ascribed to evaluating particular domains predicted the overall grades. Even though reviewers expressed greater reservations in evaluating RA exercises (vs. MDD), differences in agreement were generally small and lacking informative patterns; for example, we saw greater agreement in domain scores for RA exercises and greater agreement in strength of evidence for MDD exercises. Differences in rates of agreement by type of outcome (benefits vs. harms) were also generally small and inconsistent.

We did not find that experience doing this kind of work resulted in consist decisions. Reviewers with greater experience in conducting systematic reviews (six of eight comparisons) and those with greater experience in evaluating strength of evidence (four of eight comparisons) were more likely to agree on domain scores. In contrast, reviewers with less experience (on both measures) were more likely to agree on strength of evidence grades than reviewers with greater experience.

We found few differences in agreement based on the type of academic training of reviewers (i.e., physician or nonphysician). The intent of this exercise was to distinguish any differences in results between reviewers who would be considered clinical experts and those who would be systematic review methodologists. Because all participants were experienced reviewers from EPCs and because we did not evaluate background knowledge of the particular clinical conditions, we concluded that this distinction likely did not capture differences among reviewers in clinical and methodological expertise.

Our data were insufficient to evaluate whether the methodological approach that reviewers used to arrive at an overall strength of evidence was related to agreement on the eventual grade.

Reviewer Pairs

Agreement on strength of evidence grades across reconciled pairs, compared with agreement for independent reviewers, improved modestly, from slight to fair, across exercises that had evidence from both RCTs and observational studies. It remained fair across exercises with only RCT evidence. Final strength of evidence grades that needed to be reconciled were no more likely to be changed to a better (higher) or a worse (lower) grade. We lacked sufficient data to determine whether the mechanism used to resolve disagreements between the two independent reviewers affected the final agreed-upon grade.

Conclusions

This series of 10 exercises showed the level of diversity and complexity that EPC reviewers can encounter in their day-to-day evaluations. Our findings clearly demonstrate that the conclusions reached by experienced reviewers based on the same evidence can differ greatly. Of greatest concern is the poor level of agreement on the overall strength of evidence grades; these were typically notably lower than, and did not reflect agreement on, the domain scores from which they were intended to be derived. As presented in our analytic framework, we consider three factors that may have influenced the level of agreement on domain scores and final strength of evidence grades: reviewers' methodological approach, their judgment, and training.

Methodological Approach

By design, we included what we regarded as an "easy" exercise (Hamilton Rating Scale for Depression [HAM-D] response); it required evaluation of evidence based on meta-analysis results only from RCTs. Strength of evidence agreement for this exercise was substantial, and agreement for domain scores was high as well. This observation may indicate that the current guidance is sufficient for straightforward evaluations that can rely on quantitative tools for summarizing the available information.

In contrast, levels of agreement about overall strength of evidence across reviewers suffered when they were faced with qualitative evaluations; such groups of studies typically do not lend themselves to meta-analysis. Reviewers across EPCs used a variety of approaches that we could not clearly categorize as either methodological differences (i.e., different choice of criteria and instructions in making a decision) or judgment differences (i.e., insufficiently clear or gaps in guidance resulting in different interpretations of how to apply available guidance) and so our analysis could not distinguish which of these two categories was the likely cause of poor inter-rater reliability. Therefore, we discuss these additional concerns in relation to differences in judgment.

Judgment

Reviewers were uncertain and differed about how best to evaluate evidence that included a combination of studies with different approaches to measuring the same outcome construct. In this commonly faced problem, various measures may be considered similarly informative but differ across studies. For example, in the exercise concerning the MDD harm of sexual dysfunction, outcomes were measured as spontaneous events, abnormal ejaculation, libido decrease and different scale measures. Reviewers needed to judge how best to combine qualitatively all the data from studies with different outcome measures into domain scores.

In this example, reviewers' decisions affected scoring of three key domains: consistency, directness, and precision. EPC guidance for this step in producing a CER may not be as helpful as is needed. For example, EPC guidance states that consistency can be scored based on the similarity of outcome measures, in addition to whether selected studies differ in direction and size of effects. It says that, for directness, some, but not necessarily all, measures may be direct. Finally, it notes that precision can be evaluated with respect to whether various measures point to a clinically useful conclusion. Because the original CER authors had determined that heterogeneous measures were appropriate for inclusion in the body of evidence to evaluate this one outcome, systematic review methodologists faced a complex set of decisions and their judgments differed.

Across reviewers, combining RCT and observational studies evidence into one final strength of evidence grade was problematic and not uniformly done. Guidance directs reviewers to assess separately the domain scores for RCTs and observational studies before combining them into a final grade, but it also supports reviewers' discretion in how best to combine these separate assessments into a final grade. Reviewers can rely solely on the RCT evidence or incorporate the evidence from both RCTs and observational studies. Guidance is silent on how much weight to give observational studies relative to RCT data; in our exercises, reviewers differed in their decisions on this point.

Reviewers seemed reluctant to give a higher strength of evidence grade to findings that appeared to show no difference than to results that seemed to reflect some difference. Making this determination in these exercises was difficult because, in nine of these 10 exercises, reviewers could not rely on any quantitative assessment and because none of the exercises indicated clinically useful thresholds for precision.

To date, the EPC guidance about grading strength of evidence has reflected an orientation more toward "superiority" conclusions (and the confidence decisionmakers might have in one therapeutic option being truly better than another) than toward equivalence or "noninferiority." Said another way, the guidance is more limited when effect sizes hover around the null (i.e., of no difference).

Curiously, even for specific domain scores and strength of evidence grading where agreement was slight or fair, reviewers were generally unlikely to consider designating the assessment task to have been either difficult or very difficult. This observation may indicate that reviewers did not appreciate the complexity of the evaluations, that they relied on different, ”personal” commonly used criteria for making their decisions, or that in fact they felt reasonably confident in their decisionmaking even if the exercises seemed complicated.

Training

Inadequate reviewer training likely led to some reductions in agreement in domain scoring. All reviewers were given the EPC strength of evidence guidance chapter and it should have been the primary resource for reviewers' decisionmaking. In keeping with our goal of capturing current practice and expertise across EPCs, we did not provide reviewers with training on how to use and interpret the guidance in their evaluations. Based on reviewer comments, we believe that some reviewers did not follow clear directives in the guidance. This problem led to inappropriately considering some issues within incorrect domain categories; examples included addressing consistency concerns within risk of bias and addressing directness concerns within consistency. We saw obvious confusion about whether precision could be assessed in relation to one study, even though the guidance states that “not applicable” applies only to one study being included in assessing the consistency domain.

Experience in conducting reviews and evaluating strength of evidence translated into somewhat greater agreement on domain scores, which may reflect more experienced reviewers greater familiarity with the guidance and thus, making fewer of the mistakes listed above. However, their poorer consistency in overall strength of evidence grades, coupled with descriptions of the different approaches used across reviewer pairs, would seem to indicate that differences were not a result of inadequate training but differences in methodological approach and resulting judgments concerning how to combine scores into one final strength of evidence grade.

Study Limitations

Our findings and conclusions should be considered in light of several limitations. Because of concerns about introducing unmanageable complexity, we limited the assessment to the four “required” domains in the AHRQ EPC guidance. Thus, we did not allow reviewers to integrate findings based on dose-response association, plausible confounding, and strength of association (which are chiefly issues for observational studies and so could have aided in the integration of these studies) or publication bias.

We also required reviewers to evaluate evidence about two clinical conditions for which they may have had limited or no prior knowledge. Although we gave all reviewers background materials describing these conditions and know none worked on the original comparative effectiveness reviews, some reviewers may have had additional unmeasured prior knowledge whereas others may have been quite uninformed. We had no way of knowing who might have looked at any background materials and did not collect information that would have allowed us to control for level of knowledge in our analyses.

In addition, we had calculated and provided reviewers with p-values for exercises in which such information was missing from the original articles in the two CERs from which we took our basic information. We did not, however, calculate confidence intervals (when they were missing). Had we done so, such additional calculations might have improved agreement for, at least, the precision domain—a domain that clearly posed substantial challenges for these exercises.

Lastly, although we provided reviewers with detailed instructions on how to complete the exercises, we did not give them the criteria that the authors of the original CERs had applied to determine their own quality (or risk of bias) ratings of individual observational studies. Thus, reviewers might well have used different criteria or different scales to evaluate this particular attribute of included studies.

Needed Guidance Enhancements and Future Research

Improved Guidance for Grading Strength of Evidence

Given these findings, we conclude that additional methodological guidance, including more detailed instruction, including examples and training, is needed for the EPC program. Such guidance needs to reflect various scenarios like the range of exercises we devised, which covered variations of “thornier” bodies of evidence that EPC reviewers face with virtually every evidence report they produce. Such expanded instructions may help reviewers rely less on their, potentially idiosyncratic, individual judgments concerning appropriate domain scores than in the past and should lay out in more detail the appropriate approach for summarizing these domain scores into a strength of evidence grade.

These challenges include (although are not limited to) all of the following circumstances:

- assessing consistency, precision, and directness when dealing with outcomes that investigators have assessed through more than one measure, such as we saw with sexual dysfunction;
- assessing precision, particularly when quantitative synthesis is not possible. Reviewers need greater guidance in how to establish imprecision based on clinical thresholds compared to imprecision that is due to studies that individually may have insufficient statistical power (particularly when evaluated through p-values rather than confidence intervals);
- assessing directness when evidence for an outcome includes both intermediate and final health outcomes and may also include indirect and head-to-head comparisons;
- evaluating strength of evidence for an outcome (either benefits or harms) that requires a qualitative synthesis of the evidence, particularly if it includes a combination of RCTs and observational studies, including the relative weight to give to the different study designs and how to combine the evidence when it does not point to similar conclusions;
- scenarios when it would be appropriate to grade a body of evidence other than "insufficient" when it includes all or some study findings that are not statistically significant but the studies were not designed to be equivalence or non-inferiority trials.

We were able to provide some insights into the reasoning behind strength of evidence grades of insufficient versus moderate/high, when outcomes are being evaluated through a combination of RCTs and observational studies. However, we were unable to explain the reasoning behind reviewers' determinations of insufficient versus low strength of evidence grades, but we know that making this distinction can pose considerable challenges for those making such determinations. Thus, additional guidance will be helpful in this area as well. This is particularly true for situations in which outcomes are being evaluated through a combination of RCTs and observational studies.

Furthermore, we think that additional training for EPCs may be needed, certainly for newer or junior personnel in relation to both assessing domains and combining scores into strength of evidence grades. For example, reviewers may benefit from training in determining the appropriate domain to use to evaluate concerns (e.g., we observed at least one reviewer evaluate a consistency concern in their risk of bias assessment and overlap between consistency and precision assessments).

Reviewers bring to an evaluation varying levels of experience in making these decisions. More experienced reviewers may have strong personal preferences for how best to summarize the evidence into strength of evidence grades, based on interpretations developed by their EPC and other affiliations and they were more likely than less experienced reviewers to reach different conclusions. For these reviewers, discussions facilitated through examples, may help distinguish between differences resulting from methodology and judgment across EPCs and provide insights into where additional explanations for the reader, as well as guidance and related materials for reviewers may be needed. More detailed instructions, more use of interactive educational modules (or practice runs), and better explanations in AHRQ guidance of the rationale for arriving at different domain scores or strength of evidence grades may all be helpful.

Finally, these results raise a significant issue for transparency of EPC and AHRQ procedures. Although different approaches to this assessment task may all have validity, stakeholders need to be confident that different teams, confronted with the same evidence, would reach similar conclusions. If reviewers would apparently not reach similar grades or interpretations of that evidence, stakeholders would presumably like to know why this would happen. Improving the consistency of ultimate conclusions about evidence across EPCs (and also within EPCs) can be addressed only with consistent application of methods guidance. Related requirements and/or formats for clear explanation and documentation of how EPC review teams resolved complex decision points may be helpful. This in turn underscores our call for expanded and improved guidance about key steps, such as those noted above, for this important part of the systematic review process.

Future Research

Future research is needed to gain additional insights into reviewers' rationales for differences in domain scores and strength of evidence decisions; our study provided only a first approximation for the reasons reviewers did what they did. Additional investigations could aim to distinguish more thoroughly than we could various reasons for differences, different approaches across EPCs resulting from gaps in guidance that should be filled, areas of insufficient understanding of the guidance itself and how best to overcome that deficit, and complex decisions that may still need to be left to the review team's substantive expertise. We had limited our research to asking reviewers to explain their decisions when they had found an assessment difficult or very difficult; because relatively few indicated that this was their experience, we ended up with detailed feedback from only a small percentage of participants at several decision points.

Future reliability studies could be helpful, particularly, if they require step-by-step explanations for reviewers decisions. Testing reviewers' consistency in incorporating "additional" domains (dose-response, confounding, and strength of association), particularly when bodies of evidence include observational studies, could help to provide greater guidance and related instructional materials concerning when and how to incorporate these domains. A study could also compare whether a more standardized approach to grading strength of evidence, such as GRADE, particularly arriving at an overall grade from domain scores, would provide greater reliability than the varied approaches that EPCs have been permitted to use in the current guidance. Such work should build on the types of scenarios we tested here; of particular concern are situations when reviews include heterogeneous outcomes that do not lend themselves to statistical summarization through meta-analysis and when evidence stems from a combination of RCTs and observational studies. If inter-rater reliability is similar when methods are constrained to stipulated approaches that are less discretionary than is the case now, then such findings might indicate not only that gaps remain in both approaches but also that, for complex evaluations, no "right" approach may exist. That, in turn, highlights the importance of transparency in methodology and findings and the need for adequate documentation and explanation that speaks to the needs of all stakeholders.

References

1. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. 2008. http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed June 22, 2011.

2. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions--Agency for Healthcare Research and Quality and the Effective Health Care Program. J Clin Epidemiol. 2010 May;63(5):513-23. PMID: 19595577.

3. Atkins D, Briss PA, Eccles M, et al. Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. BMC Health Serv Res. 2005 Mar 23;5(1):25. PMID: 15788089.

4. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74. PMID: 843571.

5. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005 May;37(5):360-3. PMID: 15883903.

6. Gartlehner G, Hansen RA, Thieda P, et al. Comparative Effectiveness of Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression. Rockville, MD: Agency for Healthcare Research and Quality; Jan 2007. www.ncbi.nlm.nih.gov/books/NBK43023. PMID: 20704050

7. Donahue KE, Gartlehner G, Jonas DE, et al. Comparative Effectiveness of Drug Therapy for Rheumatoid Arthritis and Psoriatic Arthritis in Adults. Rockville, MD: Agency for Healthcare Research and Quality; Nov 2007. www.ncbi.nlm.nih.gov/books/NBK43126. PMID: 21348047.

8. Chouinard G, Saxena B, Belanger MC, et al. A Canadian multicenter, double-blind study of paroxetine and fluoxetine in major depressive disorder. J Affect Disord. 1999 Jul;54(1-2):39-48. PMID: 10403145.

9. De Wilde J, Spiers R, Mertens C, et al. A double-blind, comparative, multicentre study comparing paroxetine with fluoxetine in depressed patients. Acta Psychiatr Scand. 1993 Feb;87(2):141-5. PMID: 8447241.

10. Fava M, Amsterdam JD, Deltito JA, et al. A double-blind study of paroxetine, fluoxetine, and placebo in outpatients with major depression. Ann Clin Psychiatry. 1998 Dec;10(4):145-50. PMID: 9988054.

11. Fava M, Hoog SL, Judge RA, et al. Acute efficacy of fluoxetine versus sertraline and paroxetine in major depressive disorder including effects of baseline insomnia. J Clin Psychopharmacol. 2002 Apr;22(2):137-47. PMID: 11910258.

12. Gagiano CA. A double blind comparison of paroxetine and fluoxetine in patients with major depression. Br J Clin Res. 1993;4:145-52.

13. Schone W, Ludwig M. A double-blind study of paroxetine compared with fluoxetine in geriatric patients with major depression. J Clin Psychopharmacol. 1993 Dec;13(6 Suppl 2):34S-9S. PMID: 8106654.

14. Kroenke K, West SL, Swindle R, et al. Similar effectiveness of paroxetine, fluoxetine, and sertraline in primary care: a randomized trial. JAMA. 2001 Dec 19;286(23):2947-55. PMID: 11743835.

15. Clayton AH, Pradko JF, Croft HA, et al. Prevalence of sexual dysfunction among newer antidepressants. J Clin Psychiatry. 2002 Apr;63(4):357-66. PMID: 12000211.

16. Montejo AL, Llorca G, Izquierdo JA, et al. Incidence of sexual dysfunction associated with antidepressant agents: a prospective multicenter study of 1022 outpatients. Spanish Working Group for the Study of Psychotropic-Related Sexual Dysfunction. J Clin Psychiatry. 2001;62 Suppl 3:10-21. PMID: 11229449.

17. Jick H, Kaye JA, Jick SS. Antidepressants and the risk of suicidal behaviors. JAMA. 2004 Jul 21;292(3):338-43. PMID: 15265848.

18. Martinez C, Rietbrock S, Wise L, et al. Antidepressant treatment and the risk of fatal and non-fatal self harm in first episode depression: nested case-control study. BMJ. 2005 Feb 19;330(7488):389. PMID: 15718538.

19. Bathon JM, Martin RW, Fleischmann RM, et al. A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis. N Engl J Med. 2000 Nov 30;343(22):1586-93. PMID: 11096165.

20. Breedveld FC, Weisman MH, Kavanaugh AF, et al. The PREMIER study: A multicenter, randomized, double-blind clinical trial of combination therapy with adalimumab plus methotrexate versus methotrexate alone or adalimumab alone in patients with early, aggressive rheumatoid arthritis who had not had previous methotrexate treatment. Arthritis Rheum. 2006 Jan;54(1):26-37. PMID: 16385520.

21. Combe BG, Codreanu C, Fiocco U, et al. Double-blind comparison of Etanercept and Sulphasalazine, alone and combined, in patients with active rheumatoid arthritis despite receiving Sulphasalazine. Ann Rheum Dis. 2006 Apr 10PMID: 16606651.

22. Edwards JC, Szczepanski L, Szechinski J, et al. Efficacy of B-cell-targeted therapy with rituximab in patients with rheumatoid arthritis. N Engl J Med. 2004 Jun 17;350(25):2572-81. PMID: 15201414.

23. Klareskog L, van der Heijde D, de Jager JP, et al. Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial. Lancet. 2004 Feb 28;363(9410):675-81. PMID: 15001324.

24. Kosinski M, Kujawski SC, Martin R, et al. Health-related quality of life in early rheumatoid arthritis: impact of disease and treatment response. Am J Manag Care. 2002 Mar;8(3):231-40. PMID: 11915973.

25. van der Heijde D, Klareskog L, Boers M, et al. Comparison of different definitions to classify remission and sustained remission: 1 year TEMPO results. Ann Rheum Dis. 2005 Nov;64(11):1582-7. PMID: 15860509.

26. van der Heijde D, Klareskog L, Rodriguez-Valverde V, et al. Comparison of etanercept and methotrexate, alone and combined, in the treatment of rheumatoid arthritis: two-year clinical and radiographic results from the TEMPO study, a double-blind, randomized trial. Arthritis Rheum. 2006 Apr;54(4):1063-74. PMID: 16572441.

27. van der Heijde D, Klareskog L, Singh A, et al. Patient reported outcomes in a trial of combination therapy with etanercept and methotrexate for rheumatoid arthritis: the TEMPO trial. Ann Rheum Dis. 2006 Mar;65(3):328-34. PMID: 16079172.

28. Geborek P, Crnkic M, Petersson IF, et al. Etanercept, infliximab, and leflunomide in established rheumatoid arthritis: clinical experience using a structured follow up programme in southern Sweden. Ann Rheum Dis. 2002 Sep;61(9):793-8. PMID: 12176803.

29. Weaver AL, Lautzenheiser RL, Schiff MH, et al. Real-world effectiveness of select biologic and DMARD monotherapy and combination therapy in the treatment of rheumatoid arthritis: results from the RADIUS observational registry. Curr Med Res Opin. 2006 Jan;22(1):185-98. PMID: 16393444.

30. Listing J, Strangfeld A, Rau R, et al. Clinical and functional remission: even though biologics are superior to conventional DMARDs overall success rates remain low--results from RABBIT, the German biologics register. Arthritis Res Ther. 2006;8(3):R66. PMID: 16600016.

31. Curtis JR, Patkar N, Xie A, et al. Risk of serious bacterial infections among rheumatoid arthritis patients exposed to tumor necrosis factor alpha antagonists. Arthritis Rheum. 2007 Apr;56(4):1125-33. PMID: 17393394.

32. Schneeweiss S, Setoguchi S, Weinblatt ME, et al. Anti-tumor necrosis factor alpha therapy and the risk of serious bacterial infections in elderly patients with rheumatoid arthritis. Arthritis Rheum. 2007 Jun;56(6):1754-64. PMID: 17530704.

33. Gwet K. Handbook of inter-rater reliability: how to estimate the level of agreement between two or multiple raters. Gaithersburg, MD: STATAXIS Publishing Company; 2001.

34. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76:378-82. PMID: 5151889.

35. Feinstein AR, Chicchetti DV. High agreement but low kappa: 1. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-9. PMID: 2348207.

Appendix A. Strength of Evidence Summary Tables Used To Complete the Exercises

ANTIDEPRESSANTS: FLUOXETINE VS. PAROXETINE

1. **Benefits: HAM-D Response:**
   Key Question: For adults with MDD, do commonly used medications for depression differ in efficacy or effectiveness in treating depressive symptoms?

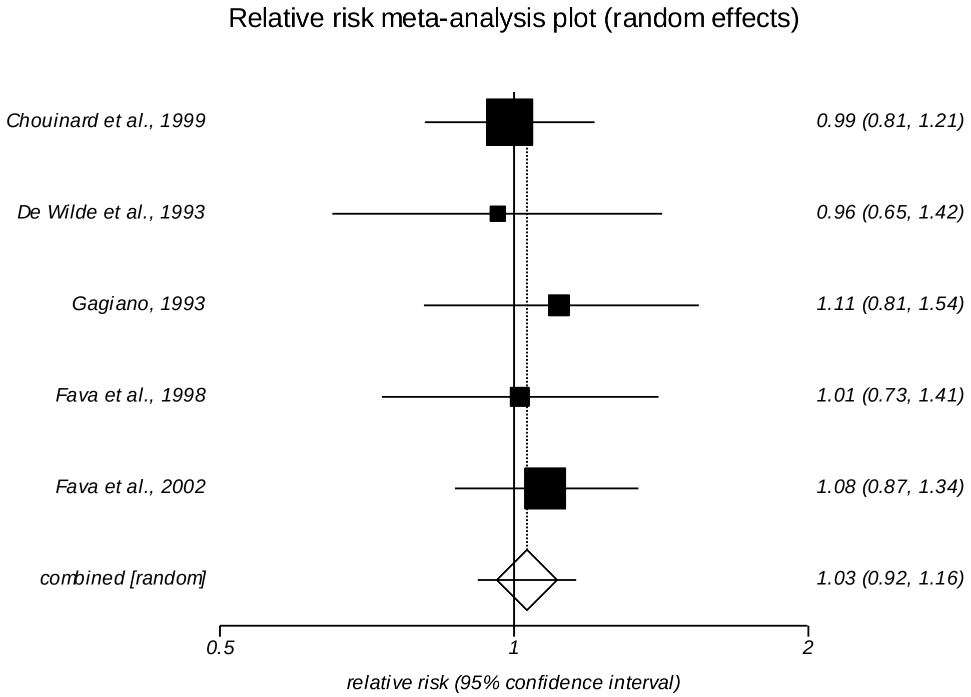
5 RCTs/TOTAL N=690

Study	N	Quality	HAM-D Responders: Fluoxetine vs. Paroxetine
De Wilde 1993[9] RCT	100	Fair	63% vs. 68% (p=NS)
Gagiano 1993[12] RCT	90	Fair	63% vs. 70% (p=NR)
Fava 1998[10] RCT	109	Fair	57% vs. 58% (p=NS)
Chouinard 1999[8] RCT	203	Fair	68% vs. 67% (p=0.93)
Fava 2002[11] RCT	188	Fair	65% vs. 69% (p=NS)

Abbreviations: NR = not reported; NS = not sufficient; RCT = randomized controlled trial = vs. = versus.

For this exercise, we also provide the pooled data analysis (forest plot on next page).

Relative risk meta-analysis of response rates comparing fluoxetine with paroxetine on the HAM-D



<u>Heterogeneity (Non-combinability) of studies</u>

Cochran Q=0.668662 (df=4) p=0.9551

Moment-based estimate of between studies variance=0

I² (inconsistency)=0% (95% CI, 0% to 64.1%)

<u>Random effects (DerSimonian-Laird)</u>

Pooled relative risk=1.031184 (95% CI, 0.918078 to 1.158226)

Chi² (test relative risk differs from 1)=0.268365 (df=1) p=0.6044

2. Benefits: Response in elderly subpopulation

Key Question: How does the efficacy of treatment with antidepressants differ in elderly or very elderly patients with MDD?

1 RCT/TOTAL N=108

Study	N	Quality	Results: Fluoxetine vs. Paroxetine
Schöne 1993[13] RCT	108	Fair	HAM-D: Significantly more paroxetine responders (Results reported in bar graph only; p=0.03) MADRS: Significantly more paroxetine responders (Results reported in bar graph only; p=0.04)

Abbreviations: HAM-D = Hamilton Rating Scale for Depression; MADRS = Montgomery Asberg Depression Rating Scale; RCT = randomized controlled trial.

3. Harms: Sexual dysfunction:

Key Question: For adults with MDD, do commonly used antidepressants differ in the occurrence of the adverse event sexual dysfunction?

4 RCTs/TOTAL N=904

2 observational studies/TOTAL N=3,154

Study	N	Quality	Results: Fluoxetine vs. Paroxetine
Fava 1998[10] RCT	109	Fair	Rate of spontaneously reported sexual dysfunction events (not explicitly defined);7% vs. 25%; (p=0.01†)
Chouinard 1999[8] RCT	203	Fair	Abnormal ejaculation: 12.2% vs. 24.3%; (p=0.16†) Impotence: 7.3% vs. 10.8%; (p=0.59†)
Fava 2002[11] RCT	188	Fair	Libido decrease: 14% vs. 15.6%; (p=0.77†) Abnormal ejaculation (corrected for gender): 11.8% vs. 20%; (p=0.34†)
Kroenke 2001[14] Open-label RCT	404	Fair	Sexual function (based on 4 individual items constituting sexual functioning scale: sexual satisfaction, ED or inadequate lubrication, difficulty having orgasm, and ability to satisfy sexual partner)—difference between groups (p=NS)
Montejo 2001[16] Prospective cohort study	487	Fair	Incidence of sexual dysfunction (assessed by Psychotropic-Related Sexual Dysfunction Questionnaire): 57.7% vs. 70.7%; (p=0.003†) Observed frequency of sexual dysfunction: Decreased libido: 50.2 vs. 63.9; (p=0.003†) Delayed orgasm/ejaculation: 49.5 vs. 63.9; (p=0.002†) Anorgasmia/no ejaculation: 39.1 vs. 52.8; (p=0.002†) Erectile dysfunction/decreased vaginal lubrication: 21.8 vs. 41.4; (p<0.0001†)
Clayton 2002[15] Cross-sectional survey	2,667	Fair	Odds of sexual dysfunction by antidepressant taken (reference drug is bupropion SR): Fluoxetine: OR, 2.23; 95% CI, 1.75 to 2.87 Paroxetine: 2.89 (95% CI, 2.24 to 3.73) Prevalence of sexual dysfunction based on CSFQ total scores lower in FLUOX-treated patients; difference reached statistical significance

†p-value calculated post-hoc by RTI.

Abbreviations: CI = confidence interval; CSFQ = Changes in Sexual Functioning Questionnaire; ED = erectile dysfunction; FLUOX = fluoxetine; NS = not sufficient; OR = odds ratio; RCT = randomized controlled trial = SR = slow release; vs. versus.

4. Harms: Suicidality

Key Question: For adults with MDD, do commonly used antidepressants differ in the occurrence of the adverse event suicidality?

1 RCT/TOTAL N=90

2 observational studies/TOTAL N=11,350

Study	N	Quality	Results: Fluoxetine vs. Paroxetine
Gagiano 1993[12] RCT	90	Fair	Suicidal ideation (HAM-D item 3) score increase: Fluoxetine: 6 (13.3%) Paroxetine: 0 (p=0.026†) Score decrease: Fluoxetine: 29 (64.4%) Paroxetine: 31 (70.5%) No patient attempted suicide
Jick 2004[17] Case-control study	1,299 nonfatal suicidal behavior cases and controls from cohort that were prescribed fluoxetine or paroxetine	Fair	Relation between fluoxetine or paroxetine and nonfatal suicidal behaviors OR (95% CI),comparing each with dothiepin as the reference group: Fluoxetine: 1.16 (95% CI, 0.91 to 1.50) [cases: 31.7%, controls: 28.5%] Paroxetine: 1.29 (95% CI, 0.97 to 1.70) [cases: 24.3%, controls: 19.4%]
Martinez 2005[18] Nested case-control study	Cases and controls, nonfatal self harm: 10,051 Cases and controls, completed suicides312:	Good	Risk of non-fatal self harm in people prescribed fluoxetine (compared with paroxetine as reference): adjusted* OR, 0.94 (95% CI, 0.79 to 1.11) Risk of completed suicides in people prescribed fluoxetine (compared with paroxetine as reference): adjusted* OR, 0.42 (95% CI, 0.13 to 1.39) *Adjusted for severity of depression, time depression was diagnosed in relation to start of txt, referral to psychiatrist or psychologist before index day, history of self harm, diagnosis of or txt for anxiety or panic disorder, schizophrenia, antipsychotic drugs, drug misuse, and alcohol misuse

†p-value calculated post-hoc by RTI

5. Harms: Nausea

Key Question: For adults with MDD, do commonly used antidepressants differ in the occurrence of the adverse event nausea?

6 RCTs/TOTAL N=689

Study	N*	Quality	Results: Fluoxetine vs. Paroxetine
De Wilde 1993[9] RCT	100	Fair	20% vs. 20.4% (p=NS)
Gagiano 1993[12] RCT	90	Fair	33.3% vs. 36.4% (p=0.764†)
Schöne 1993[13] RCT	108	Fair	11.5% vs. 9.3% (p=0.701†)
Chouinard 1999[8] RCT	203	Fair	31.7% vs. 37.3% (p=0.404†)
Fava 2002[11] RCT	188	Fair	15.2% vs. 25.0% (p=0.095†)

*N=total fluoxetine and paroxetine patients only

†p-value calculated post-hoc by RTI

ARTHRITIS DRUGS: BIOLOGICS VS. ORAL DMARDS

6. Benefits: ACR20 Response

Key Question: For patients with RA, do drug therapies differ in their ability to reduce patient-reported symptoms, to slow or limit progression of radiographic joint damage, or to maintain remission?

5 RCTs/TOTAL N=1,639

2 observational studies/TOTAL N=2,461

Study/Design	N	Comparison	Quality	Results: Biologics vs. Oral DMARDs

ERA study, 2000[19,24] RCT	424	ETN 25 mg twice/wk vs. MTX (mean 19mg/wk)	Fair	At 12 months: 72% vs. 65%, (p=0.16)
Edwards 2004[22] RCT	80	RTX vs. MTX	Fair	24 weeks: 65% vs. 38%; (p=0.025) 48 weeks: 33% vs. 20%; (p=0.204†)
TEMPO, 2005[25-27] RCT	451	ETN 25 mg twice/wk vs. MTX	Fair	At 52 weeks: 76% vs. 75%; (p=NS)
Combe 2006[21] RCT	153	ETN 25 mg twice/wk vs. SSZ	Fair	At 24 weeks: 73.8% vs. 28.0%, (p<0.01)
PREMIER, 2006[20] RCT	531	ADA 40 mg biweekly vs. MTX 20 mg/wk	Fair	Year 1: 54% vs. 63%, (p=0.043)
Geborek 2002[28] Nonrandomized open-label trial (Effectiveness trial)	269	ETN 25 mg twice/wk vs. LEF	Fair	Greater ACR20/50 responses for ETN at 3 months (p<0.001) and 6 months (p<0.05) [data reported in bar graph only]
Weaver 2006[29] Prospective cohort study	2192	ETN vs. MTX	Fair	ETN patients significantly more likely to achieve ‡mACR20 response at 12 months: adjusted* OR, 1.23 (95% CI, 1.02 to 1.47); (p<0.05) *adjusted for baseline covariates: age, baseline HAQ score (>1.5 vs. ≤1.5), comorbid disease (presence vs. absence), physician's judgment of disease severity, duration of RA, employment/disability status, Medicare coverage, race, sex, previous txt with DMARDs, previous txt with biologics, insurance status, and highest educational level attained

†p-value calculated post-hoc by RTI

‡ Excludes ESR/CRP criterion from ACR

Abbreviations: CI = confidence interval; DMARD = disease modifying antirheumatic drug; ETN = etanercept = HAQ = Health Assessment Questionnaire; mACR20 = Modified American College of Rheumatology; MTX = methotrexate; NS = not sufficient; OR = odds ratio; RA = rheumatoid arthritis; txt = treatment; vs. = versus.

7. Benefits: ACR70 Response
Key Question: For patients with RA = do drug therapies differ in their ability to reduce patient-reported symptoms = to slow or limit progression of radiographic joint damage = or to maintain remission?

5 RCTs/TOTAL N=1 =639

1 observational study/TOTAL N=269

Study/ Design	N	Comparison	Quality	Results: Biologics vs. Oral DMARDs
ERA Study = 2000[19,24] RCT	424	ETN 25 mg twice/wk vs. MTX (mean 19mg/wk)	Fair	At 12 months: no significant differences between txt groups
Edwards 2004[22] RCT	80	RTX vs. MTX	Fair	24 weeks: 15% vs. 5%; (p=NS) 48 weeks: 10% vs. 0%; (p=NS)
TEMPO = 2005[23,25-27] RCT	451	ETN 25 mg twice/wk vs. MTX	Fair	52 weeks: 24% vs. 19%; (p=0.166†)
Combe 2006[21] RCT	153	ETN 25 mg twice/wk vs. SSZ	Fair	24 weeks: 21.4% vs. 2% = (p<0.01)
PREMIER = 2006[20] RCT	531	ADA 40 mg biweekly vs. MTX 20 mg/wk	Fair	Year 1: 26% vs. 28% = (p=0.585†)
Geborek 2002[28] Nonrandomized open-label trial (Effectiveness trial)	269	ETN 25 mg twice/wk vs. LEF	Fair	12 months: differences between txt groups: (p=NS)

†p-value calculated post-hoc by RTI
Abbreviations: ETN = etanercept = LEF = lefluonomide; mg = milligram; MTX = methotrexate; NR: not reported; NS = not sufficient; RCT = randomized controlled trials; SSZ = sulfasalazine; txt = treatment; vs. = versus; wk = week.

8. Benefits: DAS Remission

Key Question: For patients with RA = do drug therapies differ in their ability to reduce patient-reported symptoms = to slow or limit progression of radiographic joint damage = or to maintain remission?

2 RCTs/N=982

1 observational study/N=1 =083

Study/Design	N	Comparison	Quality	Results: Biologics vs. Oral DMARDs
TEMPO = 2005[23,25-27] RCT	451	ETN 25 mg twice/wk vs. MTX	Fair	Remission at week 24: DAS<1.6: 13.0% vs. 13.6% (p=NS) DAS28<2.6: 13.9% vs. 13.6% (p=NS) Remission at week 52: DAS <1.6: 17.5% vs. 14% = (p=NS) DAS28<2.6: 17.5% vs. 17.1% = (p=NS)
PREMIER = 2006[20] RCT	531	ADA 40 mg biweekly vs. MTX 20 mg/wk	Fair	Clinical remission (DAS28<2.6) at 1 year: 23% vs. 21% = (p=0.582†)
Listing 2006[30] Prospective cohort study	1083	Biologics vs. conventional DMARDs	Fair	Odds of achieving remission (DAS28<2.6) at 12 months: Adjusted* OR = 1.95 (95% CI = 1.20 to 3.19); (p=0.006) *Adjusted for age = sex = # of previous DMARDs = DAS28 = ESR = FFbH = osteoporosis = previous txt with cyclosporine A. Matched pairs analysis‡ DAS28 remission at 12 months: 24.9% vs. 12.4% = (p=0.004)

† p-value calculated post-hoc by RTI

‡ Pairs of biologic and oral DMARD patients differing by less than 0.05 on propensity score

Abbreviations: ADA = adalimumab; CI = confidence interval; DAS = Disease Activity Score; DMARD = disease modifying antirheumatic drug; ESR = erythrocyte sedimentation rate; FFbH = Funktionsfragebogen Hannover Functional Status Questionnaire; ETN = etanercept; mg = milligram; MTX = methotrexate; OR = odds ratio; NS = not sufficient; RCT = randomized controlled trials; txt = treatment; vs. = versus; wk = week.

9. Harms: Serious Infections
Key Question: For patients with RA = do drug therapies differ in harms = tolerability = adherence = or adverse effects?

4 RCTs/TOTAL N=1 =215

2 observational studies/TOTAL N=7 =695

Study/Design	N	Comparison	Quality	Results: Biologics vs. Oral DMARDs
Edwards 2004[22] RCT	80	RTX vs. MTX	Fair	2 patients (5%) vs. 1 patient (2.5%) (p=NR)
TEMPO = 2005[23,25-27] RCT	451	ETN 25 mg twice/wk vs. MTX	Fair	4% vs. 4% (p=NS)
Combe 2006[21] RCT	153	ETN 25 mg twice/wk vs. SSZ	Fair	ETN: 3 serious infections in 2 patients SSZ: 0 serious infections (p=NS)
PREMIER = 2006[20] RCT	531	ADA 40 mg biweekly vs. MTX 20 mg/wk	Fair	TEAEs -# of events per 100 patient-years Any serious infection: 0.7 vs. 1.6; (p=NS) TB: 0 vs. 0
Curtis 2007[31] Retrospective cohort study	5 =326	Biologics vs. MTX	Fair	Serious infections during entire study period: 2.7% vs. 2.0%; number needed to harm=143 Crude HR (95% CI) biologic treatment association with hospitalization with a definite bacterial infection: HR = 1.39 (95% CI = 0.97 to 1.98) Adjusted* HR (95% CI) biologic treatment association with hospitalization with a definite bacterial infection: HR = 1.94 (95% CI = 1.32 to 2.83) *Adjusted for age = sex = U.S. region of residence = insurance status = comorbid diseases = corticosteroid use = MTX use
Schneeweiss 2007[32] Retrospective cohort study	2 =369	Biologics vs. MTX	Good	Compared with MTX no higher rates of serious bacterial infections in elderly patients; adjusted* model: RR = 1.0 (95% CI = 0.6 to 1.71) *Adjusted for age = sex = race = nursing home residence = hospitalization = # of physician visits = # of distinct prescription drugs = Charlson comorbidity score = RA severity = independent predictors of serious infections = previous antibiotic use = influenza vaccination and pneumococcal vaccination.

Abbreviations: CI = confidence interval; ETN = etanercept; mg = milligram; HR = hazard ratio; MTX = methotrexate; NR = not reported; NS = not sufficient; RCT = randomized controlled trials; RA = rheumatoid arthritis; RTX = TB = tuberculosis; TEAE = treatment emergent adverse event; vs. = versus; wk. = week.

10. Harms: Infusion or Injection Reaction
Key Question: For patients with RA = do drug therapies differ in harms = tolerability = adherence = or adverse effects?

4 RCTs/TOTAL N=1 =108

Study/Design	N	Comparison	Quality	Results: Biologics vs. Oral DMARDs
ERA = 2000[19,24] RCT	424	ETN 25 mg twice/wk vs. MTX (mean 19mg/wk)	Fair	37% vs. 7% (p<0.001)
Edwards 2004[22] RCT	80	RTX vs. MTX	Fair	Any event associated with first infusion: 45% vs. 30% (p=0.166†)
TEMPO = 2005[23,25-27] RCT	451	ETN 25 mg twice/wk vs. MTX	Fair	21% vs. 2% (p<0.0001)
Combe 2006[21] RCT	153	ETN vs. SSZ	Fair	32.0% vs. 2.0% (p<0.05)

†p-value calculated post-hoc by RTI

Abbreviations: ETN = etanercept; MTX = methotrexate; RTX = Rituximab; SSZ = sulfasalazine; vs. = versus; wk = week

Appendix B. Instructions Provided to Reviewers To Complete the Exercises

We provided participants with instructions and materials. The following documents were deemed *essential* materials for raters:

- Instructions for SOE reliability testing project (included in this Appendix)
- the chapter of the Methods Guide on grading the strength of a body of evidence[2]
- Background on SOE reliability testing project (included in this Appendix)
- Background on pharmacologic treatment for major depressive disorder
- Background on pharmacologic treatment for rheumatoid arthritis
  Summary table for each of the 10 grading exercises

We instructed participants to read a two-page background document that provided our project goals and approach along with a brief explanation of necessary versus supplemental information. The background document explained to participants that:

> *The project follows official guidance from AHRQ = which leaves some aspects of grading open to individual EPCs. We want to accommodate that flexibility = but we also want you to adhere to what is specified through the guidance presented in Owens (2010). Therefore = we are asking you to evaluate the four required domains for each outcome. In addition = we request that you provide separate domain scores for observational studies and RCTs. (Not all outcomes include observational studies. You will be prompted to answer accordingly for those that do.) However = we do not dictate the approach that you use to reconcile the conclusions of the two reviewers (domain scores and overall SOE grade) or the method to determine the final SOE grade for an outcome.*

We designed the exercise so that participants could ideally answer all questions by reading the background information and using the summary tables. We advised reviewers to try this approach first. If reviewers found that they wanted more information to complete the exercise = we instructed them to then consult the supplemental materials—evidence tables and full text articles—that were provided in an attached zip file.

The supplemental materials provided far more information than reviewers needed but gave them the option of looking through source documents if they so choose. In case we inadvertently overlooked including information in summary tables that individual reviewers viewed as essential based on their typical process for grading SOE = we provided full evidence tables. We also provided the full articles.

#2_INSTRUCTIONS FOR STRENGTH OF EVIDENCE RELIABILITY TESTING PROJECT

Stage I—Individual, independent grading

1. Read through this document (document name: #2-*Instructions for strength of evidence reliability testing project*); you may want to print it for easy reference as your work through the exercise.
2. Read the background document describing our overall project (document name: #3_*Background on SOE reliability testing project*).
3. Choose the condition (depression or rheumatoid arthritis) you'll address first. Read the background document for the chosen condition (#4_*Background on pharmacologic treatment for major depressive disorder*) OR (#5_*Background on pharmacologic treatment for rheumatoid arthritis*).
4. Review the document that provides the 10 outcomes you will grade for our project (document name: *SOE outcomes & corresponding summary tables*). We recommend printing this document for easy reference to the pertinent data needed to grade all outcomes. The document is divided into two sections:
   a. Antidepressants—outcomes 1-5
   b. Arthritis drugs—outcomes 6-10
5. Choose the outcome with which you would like to begin. You are not required to address the outcomes in a particular order. However, we recommend that you complete all five outcomes for one condition before moving on to the outcomes for the other condition.
6. Open the document that contains links (URLs) to the online sites for recording your responses (document name: #7_*Electronic response form links*).
7. Select the appropriate link from the table (or copy and paste the URL into your web browser) to open up the Electronic Response Form for the outcome you would like to complete first.
   a. Internet access is required to complete the Electronic Response Form
   b. For each outcome, you will need to answer the questions in the order in which they are presented.
   c. You can use the BACK button on your internet browser to go back and view completed questions while inside the Electronic Response Form.
8. At the beginning of each Electronic Response Form link, you will be prompted to type in your name. The Electronic Response Form continues with prompts for you to enter the score for each required domain (separately for RCTs and observational studies, where relevant) and the overall strength of evidence grade for the outcome. Additionally, you will be prompted to assess the difficulty encountered in scoring individual domains and overall SOE grades. Some questions offer a drop-down menu of responses; others are open-ended, offering open text boxes for you to provide spontaneous responses/assessments. You will be prompted when a text box response is required.
   a. If at any point during the completion of the Electronic Response Form you would like to take a break, select the "Save for Later Completion" button at the bottom of the screen. This will take you back to the survey you were completing. You will be taken to the first page of the survey and will need to move forward through the survey to the point at which you quit. You will notice that your previous responses have been stored.
9. At the conclusion of each Electronic Response Form, select SUBMIT to submit your responses and then close the web browser window to exit.
10. Proceed to another outcome.
11. Repeat process just described (steps 7-10).

12. While all responses are recorded/submitted electronically, you may want to keep track of your domain scores and overall SOE grades. We have provided a document (document title: #8_*Personal worksheet for SOE grading*).
    a. We recommend that you print this document. You may want to complete it by hand as you grade the 10 outcomes. You are not required to do so.
    b. RTI will send you a spreadsheet with your domain scores and overall SOE grades for all outcomes within 5 days of your completion of the exercise. The personal worksheet is only needed if you wish to see your grades across outcomes as you complete the exercise.
13. To return to a completed survey to view questions and responses, simply click on the corresponding active link
14. This completes the first stage of the SOE reliability testing exercise.

Note: A glossary of abbreviations used in the summary tables and evidence tables is attached (document name: #9_*Abbreviations*). Evidence tables and full text articles of all studies included for each outcome are provided in the attached Zip files (*MDD.zip* AND *RA.zip*).

Stage II—Paired reconciliation of scores/grades

1. Each person will be emailed a spreadsheet with the grades they submitted to RTI in Stage I.
2. Each partner group should go over each grade that they assigned and discuss any discrepancies.
3. One person in each pairing should use the Reconciliation Record document to keep track of all grades.
    a. The record should include grades that were agreed upon, as well as reconciled grades.
    b. Partner groups should take note of which domains were the most difficult, in order to report back later in the Electronic Response Form.
4. One person in each pairing will then use the Reconciliation Record document to fill out the Electronic Response Form for the Reconciliation Exercise.
    a. This form will ask for each grade to be entered and at the end, will also ask qualitative questions around the process of reconciliation.
5. To fill out the Electronic Response Form for the Reconciliation Exercise, select the link below to access the online form.
6. When prompted, fill in the name and EPC of each person in the reconciliation group.
7. Continue through the Electronic Response Form in the order the questions are presented, selecting grades from the drop-down menu and entering in responses to the qualitative questions in the text box provided.
    a. If at any time while filling out the Electronic Response Form you would like to take a break, simply select SAVE. When you re-click the link, you will be returned to that page of the online form.
8. At the end of the exercise, click SUBMIT to send the form to RTI. Close your browser window to exit.

#3_BACKGROUND ON STRENGTH OF EVIDENCE RELIABILITY TESTING PROJECT

Project Goals

This project supports AHRQ's efforts to provide guidance in methods for conducting systematic reviews (SRs) and comparative effectiveness reviews (CERs). We specifically focus on the step of grading the strength of a body of evidence (SOE). This step is described in a chapter of the AHRQ *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayProduct&productID=328) and in the Owens et al.[1] article (attached, document #10). The AHRQ education module on SOE is also available: http://www.effectivehealthcare.ahrq.gov/index.cfm/slides/?pageAction=displaySlides&tk=18.

AHRQ has a strong motivation to ensure a consistent understanding across EPCs in producing SRs and CERs; the Agency wants to promote transparency and clarity among outside stakeholders using the SRs and CERs. Thus, in designing this project, we identified several issues about SOE grading that warrant examination and created a dataset to gather empirical data from participating EPCs.

Our project focuses on reliability testing of the two main components of the task of grading SOE for specific outcomes in relation to key questions: (1) scoring evidence on the four *required* domains (risk of bias, consistency, directness, and precision) according to AHRQ's latest guidance and (2) developing an overall SOE grade given the domain scores.

The project follows official guidance from AHRQ, which leaves some aspects of grading open to individual EPCs. We want to accommodate that flexibility, but we also want you to adhere to what is specified through the guidance presented in Owens (2010). Therefore, we are asking you to evaluate the four required domains for each outcome. In addition, we request that you provide *separate* domain scores for observational studies and RCTs. (Not all outcomes include observational studies. You will be prompted to answer accordingly for those that do.) However, we do not dictate the approach that you use to reconcile the conclusions of the two reviewers (domain scores and overall SOE grade) or the method to determine the final SOE grade for an outcome.

Approach

We derived the materials for this project from two completed CERs: *Comparative Effectiveness of Drug Therapy for Rheumatoid Arthritis and Psoriatic Arthritis in Adults*;[2] and *Comparative Effectiveness of Second-generation Antidepressants in the Pharmacologic Treatment of Adult Depression.*[3] The evidence for the outcomes does not include all of the evidence included in the CERs. For our methods project, we included only a subset of the studies included in the CERs and only a subset of data from each study. Thus, the conclusions you reach will not necessarily mirror those from the CERs. No background knowledge of the comparative efficacy or harms of the drugs included in our exercise should play any role; rely solely on the evidence presented in the attached documents to determine the strength of evidence for each outcome. This is an *exercise*, not an actual CER. The sole goal of the project is to conduct a reliability test of the methods used to grade SOE.

For major depressive disorder, the drug comparison of interest is fluoxetine and paroxetine. For rheumatoid arthritis, the drug comparison of interest is between drug classes: biologics compared with oral DMARDs. The individual drug within a class should not factor into your evaluation—for the exercise, treat any individual biologic as DRUG 1 and treat any oral DMARD as DRUG 2.

A number of the included studies in our dataset include treatment arms that are not presented in the summary tables or evidence tables. Results for these arms should be disregarded. Please focus on the comparisons we have provided.

Necessary versus supplemental information: Do not be overwhelmed by the number of documents we have provided! We designed the exercise so that you could ideally answer all questions by first reading the background information and then using the summary tables. *Try this process first and see if it suffices*. If it does not, then turn to the supplemental materials—evidence tables and full text articles (provided in the attached "supplemental" zip files). We believe that we have provided far more information than you need, but you have the option of looking through source documents if you choose. We may have inadvertently overlooked including information in summary tables that you view as essential based on your typical process for grading SOE. Such information should be available to you in the evidence tables. We also provide the full articles to make available to you all relevant information. Please let us know if any tasks appear unduly burdensome or unclear before spending a lot of time trying to figure things out for yourself.

References

[1] Owens DK, Lohr KN, Atkins D, et al. Grading the strength of a body of evidence when comparing medical interventions-Agency for Healthcare Research and Quality and the Effective Health Care Program. J Clin Epidemiol. 2009 Jul 10.

[2] Donahue KE, Gartlehner G, Jonas DE, et al. (2007). Comparative effectiveness of drug therapy for rheumatoid arthritis and psoriatic arthritis in adults. AHRQ Publication No.08-EHC004-1. Rockville, MD: Agency for Healthcare Research and Quality.

[3] Gartlehner G, Hansen RA, Thieda P, et al. (2007). Comparative effectiveness of second-generation antidepressants in the pharmacologic treatment of adult depression. Comparative Effectiveness Review No. 7. (Prepared by RTI-UNC.) Rockville, MD: Agency for Healthcare Research and Quality.